Comparison of Machine Learning Algorithms for Crime Prediction in Dubai

Shaikha Khamis AlAbdouli[1], Ahmad Falah Alomosh[2], Ali Bou Nassif [3], Qassim Nasir[4]
Dept. of Sociology, University of Sharjah, Sharjah, UAE[1, 2]
Dept. of Computer Engineering, University of Sharjah, Sharjah, UAE[3]
Dept. of Computing and Informatics, University of Sharjah, Sharjah, UAE[4]

***Abstract*—This study aims to find the most accurate algorithm that is capable of predicting crimes in Dubai. It compares models on a dataset of sample crimes in the Emirate of Dubai, United Arab Emirates using the open-source data mining software WEKA, which enabled us to use Random Forest, KNN, SVM, ANN, Naïve Bayes and Decision Tree, We chose those algorithms as former studies that were effective used them. We have applied the algorithms on a dataset containing 13440 Major Crime in four categories occurred between 2014 and 2018. After comparing the models and analyzing their success rates, we identified the ideal algorithms and evaluated the effectiveness of variables in making predictions by measuring the correlation coefficients. One of the study's most crucial recommendations is to increase the variables and data, also adding more details about the crime, the criminal, and the victim. These variables make an impact on the analysis and the ultimate prediction.**

Keywords*—Machine learning; crime analysis; crime patterns; KNN; random forest; SVM; ANN; Naïve Bayes; Decision Tree; major crime

I. INTRODUCTION

The modern society in Dubai has gathered people from around all the world, more than 170 nationalities, estimated at 3,478,300 people in 2021 [1], Dubai is well known for the low crime rate [2], in Q3 of 2022, Dubai Police has reported 65% drop in the number of criminal reports at the General Department of Criminal Investigations (CID) quarterly appraisal meeting, which was presided over by Lieutenant General Abdullah Khalifa Al Marri, Commander-in-Chief of Dubai Police [3].

The rise in urban crime statistics has become a major concern for law enforcement agencies across the world. Machine learning algorithms have been progressively used to predict and prevent crime in recent years. We intend to compare the performance of various machine learning algorithms for crime prediction in Dubai in this applied scientific research. Crime is a pervasive, global social problem that lowers people's quality of life and slows economic growth [4]. As it affects people's security, crime reduction remains one of the most important social issues in large metropolitan areas [5].

In order to reduce crime by predicting and preventing it, we must have a clear understanding of the current crime situation, which requires a crime data set that enables the use of machine learning. Predicting the future occurrence of crime is more possible today than ever before with digitalization and e-governance generating data that allows for effective analysis [6].

We hope to gain insights into the most effective methods for predicting and preventing crime in this city by analyzing and comparing the accuracy, precision, and recall of these algorithms. Some research claims that crime cannot be predicted, as predictions are never 100% accurate [7]. Indeed, data is not always helpful in solving real world problems, but some scholars have succeeded in building models that helped to prevent crime [8]. This suggests that the issue with prediction may sometimes be caused by using the wrong model. Predictive policing aims to identify areas that may be subject to crimes. This is supported by routine activity theory and rational choice theory. According to both theories, a crime occurs when a person who is willing to commit it has the chance to do so and these opportunities follow patterns in both location and time rather than being distributed randomly [9].

This paper is structured as follows: Section II presents the main problem and motivation for the work. Section III presents work related to this research. Section IV describes the methodology. Section V presents the prediction models we use to analyze the data. Section VI presents our results, and Section VII summarizes our conclusions and related work.

II. PROBLEM AND MOTIVATION

There are no applied, academic studies open to students based on Dubai Crime Data as they due to the restrictions which keeps access to crime data internal and confidential.

With the unprecedented support of the Dubai Police, this applied study gave us access to real crime data in Emirates of Dubai.

By this research, we are trying to find the most accurate algorithm that is capable of predicting crimes in Dubai.

III. RELATED WORK

There are very few similar studies in the Arab Region so far. Scholars tend to conduct theoretical research and surveys, and not real crime data-based studies, On the other hand, we have found countless examples of work from other regions.

- Crime Rate Prediction Using Machine Learning and Data Mining by Sattar, Abdus and others [10] uses different clustering approaches of data mining to analyze the crime rate of Bangladesh. The authors use the KNN algorithm, and identify geographical areas that have higher crime rates, making recommendations for individuals to be cautious in those areas.

- Crime Analysis and Prediction Using Machine Learning by Olta Llaha [11] identifies the most appropriate data mining methods for analysing data collected from crime prevention sources by theoretically and practically comparing them. The authors use gender, age, employment status, and crime location as attributes. They find that data mining methods help to predict the incidence of a crime occurring and, as a result, contribute to avoiding it.

- An Experimental Study of Crime Prediction Using Machine Learning Algorithms by Sikhinam Nagamani and others [12] uses open data from Kaggle, a mix of crime types, description, time and date, and latitude and longitude to find patterns in crimes.

- Comparison of Machine Learning Algorithms for Predicting Crime Hotspots by XU ZHANG and others [13] uses an open data source from China (2015 to 2018). It suggests the use of historical crime data as well as covariates associated with criminological theories in order to evaluate the merit of machine learning algorithms.

- Crime Prediction through Urban Metrics and Statistical Learning by Luiz G. A. Alves and others [14] uses random forest regressor to predict crime and quantify the influence of urban indicators on homicides. This study finds that random forest algorithm is an excellent model for predicting crime.

- Using Machine Learning Algorithms to Analyze Crime Data by Lawrence McClendon and Natarajan Meghanathan [15] uses WEKA, open-source data mining, to conduct a comparative study between the violent crime patterns from the Communities and Crime Unnormalized Dataset provided by the University of California-Irvine repository, and actual crime statistical data for the state of Mississippi that has been provided by neighborhoodscout.com. This study finds the linear regression algorithm to be very effective and accurate in predicting crime data based on the training set input for the three algorithms.

The current study makes significant contributions by attempting to fill multiple research gaps.

First, the study adds to the relatively limited research on crime prediction in the Arab world. Our study is one of the first to use prediction models on real data from a reliable source in the Arab region.

Second, ours is one of the few research projects that has used six prediction models to determine which provides the best outcome with greater understanding and insight into the data used.

Third, to the best of the author’s knowledge (based on a search of peer-reviewed databases), no previous study has compared machine learning algorithms on crime prediction in Dubai in an applied academic setting.

IV. Methodology

A. Dataset

The crime data used in this study is confidential data individually supplied to the research team by the Dubai Police. The only publicly available crime data in Dubai are the total published by the Dubai Police, which would be insufficient for the completion of the present study [16]. This restricted and non-georeferenced dataset consists of a spread sheet that contains data compiled by the police, as shown in Table I, containing the date, the hour, the typology, used tool, the technique used, and the area of the crime, as well as the age, nationality, status and education level of the criminal for all reported crimes occurring inside the city limits between January 2014 and December 2018, amounting to approximately 52 thousand entries.

TABLE I. Data Set Used

Name	Description	Data Type
Date	Date of crime	Date
Time	Time of crime	Nominal T1: 12:00 am to 5:59am T2: 6:00am to 11:59 am T3: 12:00pm to 5:59pm T4: 6:00pm to 11:59pm
Police	Police Station responded to the crime, which refers to the area as well	String
Age	Age of criminal	Numerical
Sex	Sex of criminal	String
Nationality	Nationality of Criminal	String
Education	Educations of Criminal	String
Status	Status of Criminal	String

B. Pearson Correlation Coefficient

Pearson correlation coefficient; descriptive statistic; indicates relationship (extent of linear correlation) between two continuous variables; the better comparable the data resulting from two different methods are (i.e. the closer the correlation is) the more the r value approaches the value 1, whereby 0 represents no correlation, −1 a perfect inverse correlation (negatively sloping line) and +1 a perfect positive correlation [34].

We calculated the correlation coefficient value in order to determine how strong the association between the factors.

The correlation coefficient can be understood as follows:

- There is absolutely no association when the correlation coefficient is 0. It implies that the variables have a fully unfavorable connection. There is no association if the correlation coefficient is zero.

- If the correlation coefficient is 1, a significant positive correlation is demonstrated. It implies that the variables have their optimal positive correlation.

- A correlation coefficient with a larger absolute value denotes a stronger link between the variables.

We applied Pearson correlation in the crime dataset using Weka by selecting the attribute ranking using correlation Attribute Eval. Here are the outcomes we obtained:

TABLE II. RANKED ATTRIBUTES

Ranked Attributes	
0.1055	Nationality
0.0999	Time
0.076	Date
0.0621	Status
0.0508	Sex
0.0495	Police
0.031	Education
0.025	Age

In the Table II, we notice the most significant attribute affect for crime type is nationality with a weight = 0.105. The second largest attribute is time with a weight = 0.099. The third largest attribute is date with a weight = 0.076. The fourth largest attribute is status with weight 0.062. Next is sex, with a weight of = 0.050. Then police, with a weight of = 0.049. Finally, the last two attributes are education with a weight of = 0.031, and age with a weight of = 0.025.

C. Preprocessing

First, we selected four major crime typologies out of 10. The Major Crimes are categorized as: (Willful Murder, Aggravated Assault, Rape, Robbery, Theft, Abduction, Grand Auto Theft, Burglary, Drugs, Human Trafficking) due to the Non-disclosure Agreement we cannot declare which four categories we have chosen. We removed any crimes that had missing values due to missing data or compiling errors. This reduced the number of entries to 13,440.

Instead of using exact times, we categorized hours into four periods, 6am to 12pm, 12pm to 6pm, 6pm to 12am, and 12am to 6am. We categorized nationalities into three groups: Gulf Countries (Saudi Arabia, United Arab Emirates, Oman, Kuwait, Bahrain, Qatar) , Arab countries (Algeria, Comoros, Djibouti, Egypt, Iraq, Jordan, Lebanon, Libya, Mauritania, Morocco, Palestine, Somalia, Sudan, Syria, Tunisia and Yemen) , and rest of world.

To train and validate the data, the dataset is divided into various subsets with 10 folds in cross-validation, the training was on 70% and test was on 30%.

D. Evaluation Metrics

*1) Accuracy:*The percentage of overall predictions that were correct.

2) Accuracy: ((TP + TN) / (TP + FP + TN + FN)) * 100

3) Precision: Precision reveals the proportion of genuinely positive forecasts among all positive ones. The ratio of accurately positive predictions to all positive predictions is how it is defined.

4) Precision = Predictions accurately positive / Total predicted.

5) Precision = TP/TP+FP

6) Recall: Shows how many truly positive values were predicted out of all positive values. It measures the proportion of accurate positive predictions to all the positive examples found in the dataset.

7) Recall: It is the ratio of predicted values that came true to actual values in the dataset.

8) Recall = TP/TP+FN

9) F1 Score: It is the harmonic mean for precision and recall values as depicted in Fig. 1. [17]

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

Fig. 1. F1 equation.

V. PREDICTION MODELS

A. Random Forest

During the training phase of the random forests or random decision forests ensemble learning approach (which is used for classification, regression, and other tasks), a large number of Decision Trees are built. For classification problems, the random forest output is the class that the majority of the trees choose. For regression tasks, the mean or average prediction of each individual tree is returned. Random decision forests correct Decision Trees' proclivity for overfitting their training dataset [18] [19] [20] [21].

The classifier used is Random Forest.

B. (The K-Nearest Neighbor's Algorithm) KNN

The K-Nearest-Neighbours (KNN) is a non-parametric classification method, which is simple but effective in many cases [22]. KNN is a non-parametric classification algorithm, it works as a supervised learning algorithm. A labeled training dataset is provided where the data points are categorized into various classes, so that class of the unlabeled data can be predicted [23].

The classifier used in KNN is IBk.

C. Support-Vector Machines (SVM)

Support vector machines (SVMs) can be used to handle classification, regression, and outlier problems that are frequently encountered in supervised learning [24].

The mathematical pedigree of SVMs is the best of any statistical learning procedure. It was created as a classifier that maximizes a slightly different definition of a margin, resulting in a novel "hinge" loss function [25]. Weka can classify objects using the support vector machines algorithm [26].

D. Artificial Neural Networks (ANNs)

Artificial Neural Networks can be defined as systems designed to model functions that simulate the human brain [27]. They are increasingly being used to model complex, nonlinear phenomena [28]. ANNs are nonlinear, adaptive information processing systems that are made up of many interconnected processing units. ANNs have functions such as associative memory, nonlinear mapping, classification recognition, and optimization computation as an effective empirical modelling tool [29].

The classifier used in ANN is Multilayer Perception.

E. Naive Bayes

The naive Bayes classifier significantly simplify mastering through assuming that capabilities are impartial given class [30]. Naive Bayes is a probability classification model that makes machine learning easier by performing calculations on datasets with the goal of predicting probabilities in a class under the assumption of strong independence. Classification is a type of directed learning [31].

The classifier used in Naive Bayes is Naive Bayes

F. Decision Tree

Decision tree is one of the popular predictive modelling approaches used in many areas including statistics, data mining and machine learning [32]. Decision tree classifiers are regarded to be a standout of the most well-known methods to data classification representation of classifiers [33].

The classifier used in Decision Tree is J48 which is a Decision Tree classification algorithm based on Iterative Dichotomiser 3.

VI. Results

We can summarize the results in Table III:

TABLE III. ALGORITHMS RESULTS

Algorithm	Accuracy	Recall	Precision	F1
Random forest	76.986%	0.77	0.77	0.769
KNN	78.474%	0.785	0.789	0.784
SVM	54.575%	0.546	0.524	0.520
ANN	51.093%	0.511	0.511	0.509
Naïve Bayes	53.526%	0.535	0.516	0.511
Decision Tree (J48)	67.976%	0.680	0.678	0.678

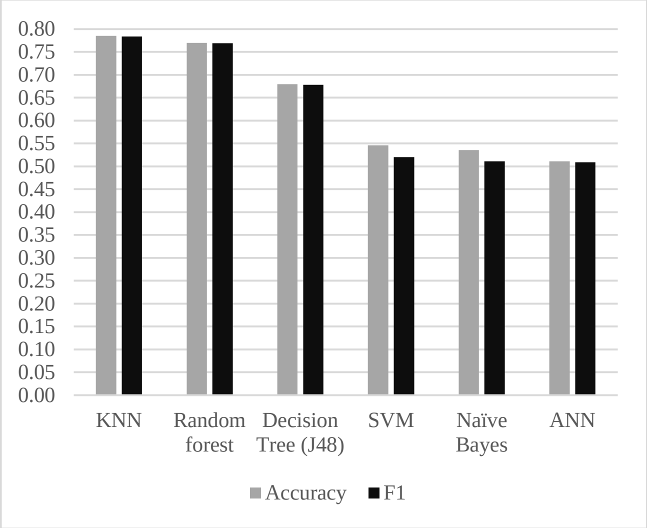

Fig. 2. Accuaracy and F1 results.

In the preceding Table III and Fig. 2, we observe that Random Forest and KNN achieve the best results. KNN achieves the best results where accuracy = 78.474%, and F1 = 0.784. Next comes Random Forest with accuracy = 76.986%, and F1= 0.769. Decision Tree achieves good results with accuracy = 67.976%, and F1 = 0.678. SVM, Naïve Bayes, and ANN achieve low performance. SVM achieves accuracy = 54.58%, and F1 = 0.520. Then comes Naïve Bayes with low results of accuracy = 53.526%, and F1 = 0.511. Last comes ANN with the lowest results: accuracy = 51.093%, and F1 = 0.509. For time complexity, Random Forest, KNN, Naïve Bayes, and Decision Tree take seconds while SVM takes minutes. ANN takes more than two hours.

VII. Conclusions and Future Work

This study compared several popular machine learning algorithms for use in crime prediction, including KNN, Random Forest, SVM, ANN, Nave Bayes, and Decision Tree.

Our findings show that these algorithms can provide useful insights into predicting crime patterns, with KNN having the highest overall accuracy (78.474%) and F1 scores. The performance of each algorithm, however, varied depending on the dataset and crime type being analyzed.

According to our findings, using machine learning for crime prediction has the potential to improve public safety and law enforcement efforts. However, it is critical to recognize the limitations and ethical concerns associated with the use of predictive algorithms in criminal justice systems. Because machine learning models are only as good as the data on which they are trained, it is critical to ensure that crime prediction datasets are diverse and representative of the population. Furthermore, it is critical to address potential biases and avoid discrimination when deploying these models.

Also, by using the correlation, we discovered that adding more attributes, and detaching and elaborating the current data rather than grouping it into periods, may yield better results in the future.

Overall, our research highlights the potential and challenges of using machine learning to predict crime in Dubai. As the field develops, it will be critical to carefully evaluate and refine these algorithms to ensure their accuracy, fairness, and ethical implementation.

We suggest that future work in this area include more variables, such as: data about buildings, street names, exact locations containing longitude and latitude, data about the victims, income, and the relationship between the criminal and victim.

References

[1] Population and Vital Statistics. Dubai Statistics Center. (n.d.). Retrieved January 22, 2023, from https://www.dsc.gov.ae/en-us/Themes/Pages/Population-and-Vital-Statistics.aspx?Theme=42

[2] Police, D. (2023, January 13). Major Crime Statistics. Dubai Police . Retrieved January 15, 2023, from https://www.dubaipolice.gov.ae/wps/portal/home/opendata/majorcrimestatistics

[3] WAM. (2022, October 15). Dubai Police records 65% drop in criminal reports during Q3. Wam. https://www.wam.ae/en/details/1395303092140

[4] Bappee, F.K., Soares, A., Petry, L.M. et al. Examining the impact of cross-domain learning on crime prediction. J Big Data 8, 96 (2021). https://doi.org/10.1186/s40537-021-00489-9

[5] Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. DOI:10.1007/978-981-15-7394-1_5

[6] Lenin Mookiah||William Eberle||Ambareen Siraj (2015). Survey of Crime Analysis and Prediction. Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2015).

[7] Sathyadevan, Shiju & S., Devan & Gangadharan, Surya. (2014). Crime Analysis and Prediction Using Data Mining. 10.1109/CNSC.2014.6906719.

[8] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[9] Stalidis, Panagiotis & Semertzidis, Theodoros & Daras, Petros. (2021). Examining Deep Learning Architectures for Crime Classification and Prediction. Forecasting. 3. 741-762. 10.3390/forecast3040046.-

[10] Sattar, Abdus. (2021). Crime Rate Prediction Using Machine Learning and Data Mining. DOI:10.1007/978-981-15-7394-1_5

[11] O. Llaha, "Crime Analysis and Prediction using Machine Learning," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2020, pp. 496-501, doi: 10.23919/MIPRO48935.2020.9245120.

[12] Nagamani, Sikhinam & Bhavishya, & Kumar, Mr & Sree, T & Reddy, Lakireddy. (2022). An experimental study of Crime Prediction using Machine Learning Algorithms. Test Engineering and Management. 83. 17819 - 17825.

[13] X. Zhang, L. Liu, L. Xiao and J. Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots," in IEEE Access, vol. 8, pp. 181302-181310, 2020, doi: 10.1109/ACCESS.2020.3028420.

[14] Alves, Luiz & Valentin Ribeiro, Haroldo & Rodrigues, Francisco. (2017). Crime prediction through urban metrics and statistical learning. https://doi.org/10.1016/j.physa.2018.03.084

[15] McClendon, Lawrence & Meghanathan, Natarajan. (2015). Using Machine Learning Algorithms to Analyze Crime Data. Machine Learning and Applications: An International Journal. 2. 1-12. 10.5121/mlaij.2015.2101.

[16] Major Crime Statistics. (n.d.). Dubai Police. Retrieved February 8, 2023, from https://www.dubaipolice.gov.ae/wps/portal/home/opendata/majorcrimestatistics

[17] Narain, Profbhavana. (2021). An Empirical Analysis of Machine Learning Algorithms For Crime Prediction using Stacked Generalization: An Ensemble Approach. IEEE Access. XX. 1-9. 10.1109/ACCESS.2021.3075140,.

[18] Azhari, Mourad & Alaoui, Altaf & Acharoui, Zakia & Ettaki, Badia & Zerouaoui, Jamal. (2019). Adaptation of the random forest method: solving the problem of pulsar search. SCA '19: Proceedings of the 4th International Conference on Smart City Applications. 1-6. 10.1145/3368756.3369004.

[19] Cutler, Adele & Cutler, David & Stevens, John. (2011). Random Forests. 10.1007/978-1-4419-9326-7_5.

[20] Ali, Jehad & Khan, Rehanullah & Ahmad, Nasir & Maqsood, Imran. (2012). Random Forests and Decision Trees. International Journal of Computer Science Issues(IJCSI). 9.

[21] Oshiro, Thais & Perez, Pedro & Baranauskas, José. (2012). How Many Trees in a Random Forest?. Lecture notes in computer science. 7376. 10.1007/978-3-642-31537-4_13.

[22] Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin. (2004). KNN Model-Based Approach in Classification.

[23] Taunk, Kashvi & De, Sanjukta & Verma, Srishti & Swetapadma, Aleena. (2019). A Brief Review of Nearest Neighbor Algorithm for Learning and Classification.1255-1260.10.1109/ICCS45141.2019.9065747.

[24] Md Imran Hossain. (2022). Support Vector Machine*. Frankfurt University of Applied Sciences. Frankfurt. Research for Master of Science in High Integrity Systems.

[25] Berk, Richard. (2020). Support Vector Machines. 10.1007/978-3-030-40189-4_7.

[26] Bell, Jason. (2015). Support Vector Machines. 10.1002/9781119183464.ch7

[27] Akgül, İsmail & Kaya, Volkan. (2022). A REVIEW ON ARTIFICIAL NEURAL NETWORKS.https://www.researchgate.net/publication/360967369_A_REVIEW_ON_ARTIFICIAL_NEURAL_NETWORKS

[28] Yang, X.. (2009). Artificial neural networks. Handbook of Research on Geoinformatics. 122-128. 10.4018/978-1-59140-995-3.ch016.

[29] Wang, Haonan & Chen, Yijia. (2022). Application of Artificial Neural Networks in Chemical Process Control. Asian Journal of Research in Computer Science. 22-37. 10.9734/ajrcos/2022/v14i130325.

[30] Sai, Mitra & Kamasani, Sai Mitra. (2021). A STUDY ON NAIVE BAYES CLASSIFIER. https://www.researchgate.net/publication/356267142_A_STUDY_ON_NAIVE_BAYES_CLASSIFIER

[31] Afdhaluzzikri, Afdhaluzzikri & Mawengkang, Herman & Sitompul, Opim. (2022). Perfomance analysis of Naive Bayes method with data weighting. SinkrOn. 7. 817-821. 10.33395/sinkron.v7i3.11516.

[32] Bshouty, Nader & Haddad-Zaknoon, Catherine. (2021). On Learning and Testing Decision Tree. https://doi.org/10.48550/arXiv.2108.04587

[33] Jijo, Bahzad & Mohsin Abdulazeez, Adnan. (2021). Classification Based on Decision Tree Algorithm for Machine Learning. Journal of Applied Science and Technology Trends. 2. 20-28.

[34] Nahler, Gerhard. (2010). Pearson correlation coefficient. 10.1007/978-3-211-89836-9_1025.