THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 4. Issue 1 (2024)
https://doi.org/ 10.62019/abbdm.v4i1.76

Predicting Crime rate in Punjab: An Analysis of Artificial Intelligence based Models

Bisma Ali*

Chronicle

Article history
Received: January 1st, 2024
Received in the **revised format:** Feb 29, 2024
Accepted: March 21, 2024
Available online: March 28, 2024

Bisma Ali is a student of Lahore Grammar School, 55 Main Gulberg Lahore, Pakistan.
Email: bismaalirajalgs@gmail.com

Abstract

This study aims at measuring and evaluating the accuracy rate or Mean Square Error (MSE) of different machine learning models to predict crime count (dependent variable) on the basis of a number of independent variables such as shift, day, month, police station, district and time of the day. It is a quantitative research study where primary data is collected from rescue 15 call record. The year 2022 has been chosen for the purpose of data under the crime head of theft across the Punjab province which is divided into three eight hourly shifts each in a day during any week of a month. The crime count indicates number of incidents of theft reported during these three shifts per day per month in a district in 2022. The data containing variables is then used for prediction through an appropriate machine learning model. The MSE value of Random Forest Regression (RFR) is greater than MSE value of Multi-Layer Perceptron Regression (MLPR). It is however lower than MSE score of Linear Regression (LR) model and greater than MSE value of Support Vector Regression (SVR). This shows that MLPR model is the most suitable one with least error as compared to other models which have higher MSE values while predicting the dependent variable (crime count). The study findings may assist the intelligence branches of police to efficiently utilize their material and human resources while applying the appropriate model of machine learning to predict the number of crime incidents in any geographical dispensation at sub national level.

***Corresponding Author:**

Keywords: Machine Learning, Algorithm, Artificial Intelligence, MLP Regression, SVR model, LR model, RFR model.

INTRODUCTION

Predicting crime pattern is an essential requirement of modern policing. It helps in assessing the future needs of police and how or where to utilize its resources in an efficient and effective way. The predictive crime patterns are obtained through available data on crime, violations, arrests, interviews and photographs. Nowadays data bank of DNAs has been effectively used to determine prospective criminals involved in an incident in a particular area with the assistance of the artificial intelligence. As the crime, both property and person, is influenced by increase in population, fluctuation in economic conditions, population migration, urban area expansion etc therefore police department may get immense support from the artificial intelligence in pointing out the areas where (i) new policing structures need to be evolved, (ii) enhanced financial mobilizations are to be made and (iii) human resources must be spent to monitor and detect the crime in coming days. The crime pattern in Punjab is not uniform. It varies from district to district and from region to region as well. It is also seasonal in nature with certain criminal activities shooting up in winters as compared to in summers. It is also even pegged with

special occasions such as Eid breaks or harvesting season when workforce comes back to their native towns, cities and villages the crime against property increases appreciably. Moreover, the 36 districts in the province have their own particular features as well. As we move from South towards the central region the highway robberies and snatching incidents increase. In South where Sind, Balochistan and Khyber Pakhtunkwa (KP) boundaries converge with each other a different nature of criminal activities are present. Gangs of armed men in Kacha area of Rajanpur and RY Khan are operating who often honey trapped or kidnapped individuals for ransom and take them to the Sindh province only to be released after ransom through local intermediaries. The provincial capital has a daily flux of population moving into and outside the city causing vast avenues for the criminals to commit crime. The north part of the province abutting Khyber Pakhtunkwa (KP) exhibits a kaleidoscope of criminal activities in which individuals from KP play a major role. Rawalpindi district, in the vicinity of Islamabad, the capital of Pakistan, has assumed a sensitivity benchmark owing to presence of considerable military establishments and existence of diplomatic enclave. This makes it essential that an AI based predicting model should be present to guide local police formations to focus upon particular areas of crime and its detection. The contours of such a model can assist the police commanders in districts, city districts or capital city district Lahore to invest their resources in field efficiently and maintain a high efficiency of output as well.

The 15-call system in Punjab has been automated and integrated in 2022. By virtue of its connectivity each and every call across the province is digitally centralized. The process involved in designing a predictive model is that data sets have been created on account of 15 calls received at Safe City Lahore. A particular sub set of these calls pertaining to crime against property has been examined. Nearly 38971 calls have been recorded pertaining to theft throughout Punjab. Three shifts, eight hours each, have been earmarked as A, B and C respectively. The day and crime count have also been registered. The crime count indicates number of times a crime of similar nature has taken place. For example, in Attock, 1093 incidents have been reported during the week days in which crime count was calculated as 2024 during various shifts of the day. However, the crime count varies in each shift of each day in a month. In January 2022, 420 count was calculated in the first shift, 564 during second shift and 135 in the third shift respectively. Therefore, the variables used in the current study include (i) day (ii) month (ii) shift (iv) crime count and (v) district. A sample of these variables pertaining to district Attock during the month of January 2022 has been indicated below

Table 1.
Crime counts along with district, month, and days

District	Month	Crime Count			Total Crime Count	Days
		A	B	C		
Attock	January	66	73	24	163	Saturday
		69	105	23	197	Sunday
		65	74	15	154	Monday
		54	77	22	153	Tuesday
		47	74	15	136	Wednesday
		51	72	17	140	Thursday
		68	89	19	176	Friday
Total		**420**	**564**	**135**	**1119**	

Data Accumulation and Model Training

The data processing is a mechanism of information processing through observation and examination. A number of steps are involved in data processing which are known as DP functions. The validation of the available data is essential since it ensures which data is correct and relevant. Sometimes correct data is obtained without any relevancy to the study. Similarly, a relevant incorrect data may also lead to an incorrect predictive modeling. Once the validation of data has been completed the items within the data are sorted under different heads or sub heads in the current scenario. We have chosen eight different heads under which sorting has been done. These include year, month, day, district, duty shift, crime count and day of a week. During sorting summarization of detailed data into its main points is also carried out. This involves statistical methods examination or automatic tools available with different software applications. Once summarization has been completed, the process of combining multiple pieces of data is done.

The aggregate data is used for multiple reasons especially recognizing trends and patterns of processes to gain relevant insight and assess current measures for data modeling. The data is then interpreted and presented for analysis. A detailed summary of such an examination is made under various categories or heads as a whole. uring data processing all irrelevant data is taken out and all variables representing important data are kept for modeling. This is also known as data cleaning and transmission. It also involves removing typographical errors or validating or correcting values against a known list of entries under a head or sub heads. Data harmonization is also included in it in which varying file format, naming convictions and columns are transformed into a cohesive set of data. The AI model is a tool or algorithm which is based on a certain data set through which it can arrive at a decision without any external assistance or interference. The model recognizes certain patterns on the basis of available data set which allows it to reach at a certain conclusion to make a prediction on the basis of available information and data. These models include Linear Regression, Support Vector Regression (SVR), Multilayer Perceptron Model (MLP) and Random Forest Model respectively. Once these models have been selected the next step is to train these models on the available data. It is a process of teaching an AI system to perceive the available data so that it can read it properly without any error or inaccuracy to develop a significant ability to interpret on basis of its learning.

Once it develops the capability of perceiving, interpreting and learning the data it becomes inherently capable of inferencing thus, making decisions based on the input data. Usually, huge amount of data is fed into the system and any output by it is analyzed and examined. Certain adjustments are also made based on the output accuracy of the model. A validation and testing are also carried out later on to see whether the system made accurate decisions leading to passing or failing a test. If it fails, the model is retrained, readjusted and similar process is repeated again and again till it makes accurate decisions. There are three types of training of AI model such as supervised, reinforcement and unsupervised training. Here we stick to supervised learning of the four different models on basis of available data. The data testing is not one time activity but is performed continuously. Once each and every model have been trained or tested, most appropriate model is chosen. It is a trial-and-error process in which some hyper parameters are run on the algorithm and compare its performance on a validation set. Model can be tuned manually or automatedly. After tuning a model, a prediction can be made with certain percentage of error.

LITERATURE REVIEW

Predicting geographical crime information is an effective use of machine learning techniques. Thus, the Support Vector Machine (SVM) method was used in 2006 to forecast crime locations in Columbus, Ohio, in the United States. In order to forecast the hot spot region and increase its efficacy, SVM employed both random and clustering algorithms to the training and test datasets (Kianmehr & Alhajj, 2006). These algorithms are used to investigate the relationship between the motivations behind crimes and their incidence. In order to predict the relationship between burglaries and a number of other factors, including time of day, day of the week, barriers, connectors, and repeat victimization, a Logistic Regression (LR) algorithm was implemented in 2013. However, this model proved to be unsuccessful for a large geographic area (Antolos et al., 2013).After employing the SmoteR algorithm to identify more serious crimes, the Random Forest (RF) approach was used in 2015 to forecast crime in the southern US states.

Furthermore, when the population and density were chosen as actual numbers, their work was optimized using R software (Cavadas et al., 2015). In the end, the auto-regressive technique was used to estimate the quantity of crimes that occurred simultaneously and in metropolitan locations (Cesario et al., 2016). The Naive Bayes (NB) algorithm was introduced in 2017 and is designed to forecast crime incidents based on historical data demonstrating similar crimes occurring in the same location.An investigation carried out in India focuses on the many categories of crimes and how often they occur in various locations and periods. A thorough study is conducted with murder being the most common sort of crime compared to the other categories. It is found that the scaled method produced the best result for the considered data when compared to the other two using the Bayesian, Levenberg, and Scaled algorithms on both train and test data. An iteration is also calculated at which highest valid performance was reached. According to the study, there is a possibility of reducing the crime rate to 78 percent, which suggests an accuracy of 78 (Shraddha & Vijayalakshmi, 2020).

The authors suggest using data-driven analysis to extract meaningful information from Indian crime statistics. Police and other law enforcement agencies in India may find the suggested strategy useful in managing and preventing crime on a regional basis. The suggested method preprocesses the data using MySQL Workbench and R programming, then builds several regression models based on various regression algorithms, namely random forest regression (RFR), decision tree regression (DTR), multiple linear regression (MLR), simple linear regression (SLR), and support vector regression (SVR). When given the appropriate inputs, these regression models can predict 28 distinct categories of Indian Penal Code (IPC) cognizable crime counts as well as the overall number of IPC cognizable crime counts by area, state, and year (for the entire nation). Moreover, data which has been pre-processed (corresponding to the years 2014 to 2020) and data that has been forecasted by the comparatively best regression model for the year 2022 are visualized using data visualization techniques, namely chord diagrams and map plots. With an adjusted R squared value of 0.96 and a MAPE value of 0.2, Random Forest Regression (RFR), which predicts total IPC cognizable crime, is found to fit the data the best. Among regression models that predict theft crime counts region-wise, the random forest regression-based model fits the data the best, with an adjusted R squared value of 0.96 and a MAPE value of 0.166. According to these regression models, the state of

Andhra Pradesh is expected to have the greatest crime rates, with the highest anticipated crime rate of 31,933 being in the Adilabad district (Aziz et al., 2022). Bangladeshi crime trends and patterns are predicted using three machine learning models. By using certain novel measures, this analysis may assist the Bangladeshi police department and several other law enforcement organizations in reducing crime. Various regression models are trained using historical crime data. Regression models are used to forecast for the year 2018 when training is finished. It has been noted that for this specific dataset, random forest and polynomial regressions perform better when making predictions. Random forest regression, on the other hand, fared better than linear regression; nevertheless, it does not scale very well when it comes to the training set for time-series data. The actual crime data from prior years is gathered for this study from the Bangladesh police website. The dataset includes total counts of various crimes that have been classified by Bangladesh's police force. The collection is categorized into two groups such as divisional region data and metropolitan region data, each representing a different region of Bangladesh.

The 840 incidents or crimes that make up the dataset were gathered from all throughout the nation. One objective feature and three prediction features make up the dataset. The area, month, and year are the three predictive factors. The anticipated value of various crime categories is the aim feature. Each instance's data comes from a different Bangladeshi area. The data mining approach is employed to predict Bangladesh's future crime patterns. The linear regression model is trained using past year's crime data for this purpose. Following the training of linear regression, many categories of criminal activity are predicted for 2016. The experimental results also show that most crimes are rising in tandem with population expansion. Therefore, the information gleaned from crime data analysis may help the police department and other law enforcement organizations predict, stop, or resolve Bangladesh's future crime patterns (Biswas & Basak, 2019). Bangladesh's high crime rate is a result of several socioeconomic problems, including population expansion and poverty. Understanding crime trends is crucial for law enforcement organizations to stop criminal conduct in the future. These authorities require a structured crime database for this reason. In this study, a unique crime dataset containing demographic, meteorological, spatial, and temporal data regarding 6574 crime occurrences in Bangladesh is introduced.

Then, using this freshly constructed dataset, five supervised machine learning classification methods are assessed, yielding good results. Additionally, exploratory analysis was carried out on several facets of the dataset. It is anticipated that crime incidence prediction systems for Bangladesh and other nations would be built around this information. Law enforcement authorities can get benefit from the study's findings in forecasting and containing crime as well as ensuring the best possible resource allocation for crime patrol and prevention (Shohan et al., 2022). The 'City of Chicago from 2013 to 2017' dataset was used to forecast different categories of crimes. In this investigation, Decision Tree and Naïve Bayes approaches were employed. In the preprocessing phase, listwise deletion was used to manage missing data. The backward feature selection approach was used for feature selection. The Decision Tree classifier outperformed Naïve Bayes in terms of performance, with a prediction accuracy of 91.59%. The accuracy of Naive Bayes, however, is 83.40%. Decision Tree therefore fared better than Naïve Bayes when comparing the two findings (Aldossari et al., 2020). Measuring and evaluating the

accuracy rate or mean square error of different machine learning models to predict the target variable (crime count) in presence of the feature variables (day, week, month, duty shift, year). Measuring and evaluating the accuracy rate or mean square error of different machine learning models to predict the target variable (crime count) in presence of the feature variables (day, week, month, duty shift, year). Which model is best to predict accurately the target variable (crime count) and feature variables (District, Duty shift, Month, Day of week)?

RESEARCH SIGNIFICANCE

Governments can make better judgments with the help of machine learning. Forecasts can inform improved decision-making, and machine learning aids regarding crime prediction while highlighting irregularities in legal systems. Reliable forecasts of criminal activity in the future can benefit the economy and society. The employment of crime detection systems in society might have a similar effect on violent crime by warning potential offenders of impending crimes and increasing the possibility that they would be apprehended. Sustainable development and economic prosperity are correlated with safer and more secure environments. Predicting crimes before they happen helps prevent property losses and human casualties. Understanding the nature of a crime is crucial for creating a highly effective crime prediction model (Elluri et al. 2019). Features related to the crime, such as the victim(s)' age, gender, location, economic status, education status, time, date, day of the week, year, and month, as well as the offender(s)' age, gender, and number of offenses, income, and weapon used, can all be considered as various aspects of the crime's nature. The rate of crime in Punjab is rising daily and has emerged as one of the most difficult issues. A system that is capable of identifying and forecasting these actions is required. Though a lot of models have previously been created to lower crime rates, more work has to be done to increase the precision and accuracy of these models. When it comes to predicting the location, crime rate, and timing of crimes, researchers and government security officers encounter certain challenges. They also have difficulties in selecting an efficient approach. By addressing the issue of projecting the number of crimes by year, day of the week, and duty shift in the context of Pakistan, this study closes the gap. This study is distinct from others since it examines aspects and factors that haven't been examined in previous research. Crime is unpredictable and is unpredictably occurring every day. To ascertain if the crime rate has increased or decreased from previous years, crime forecast is important. The goal of employing machine learning and data mining for crime detection is to lower the levels of crime.

METHODOLOGY

Primary data for this quantitative research study was gathered from the rescue 15 call record. For the aim of gathering data under the criminal head of theft, the year 2022 has been selected. The Punjab province is divided into three eight hourly shifts (A, B, and C) each day on any weekday of the month. The crime tally shows how many theft events were recorded in a district in 2022 during these three shifts daily in a month. The factors in the data are utilized by a suitable machine learning model to make predictions. The filtered data are gathered in order to compute the accuracy. After that, it is separated into data sets for training and assessment. There are several machine learning algorithms that may be used to forecast the number of crimes. To determine the best fitted model

with the maximum accuracy, supervised learning approaches such as the Support Vector Machine, Linear Regression model, MLP Regressor, and Random Forest Regressor model have been studied.

RESULTS

Dataset is uploaded on Jupiter notebook where below following coding is applied to fulfil the data preprocessing steps and implementing the machine learning models.

```
import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns from datetime

import datetime from sklearn.preprocessing

import MinMaxScaler
```

Pandas is a popular open-source data manipulation and analysis library for Python. It provides easy-to-use data structures (primarily, the DataFrame and Series) and functions designed to efficiently manipulate large datasets. **matplotlib.pyplot** is a collection of functions in the popular Python plotting library, Matplotlib. Matplotlib is a 2D plotting library that produces static, animated, and interactive visualizations in Python. **Seaborn** is a Python data visualization library based on Matplotlib. It provides a high-level interface for creating informative and attractive statistical graphics. The statement from **datetime import datetime** is a Python import statement that brings the datetime class from the datetime module into the current namespace. The **MinMaxScaler** is a feature scaling technique provided by scikit-learn (sklearn). Feature scaling is a crucial step in many machine learning algorithms to ensure that numeric features have the same scale.

```
df pd.read_excel('pucar2022v2.xlsx')

print(df.head())
```

Data from excel value is extracted or uploaded by using **df function**. After loading the file data on jupiyter platform, **df.head function** is applied to show the heading of data like Year, Month, Day, District, Dutyshift, crimecount, day of week.

```
filtered_dfdf [df['District'] != 'Lahore']
```

The statement filtered_df = df[df['District'] != 'Lahore'] is a pandas DataFrame operation that filters rows based on a condition and assigns the result to a new DataFrame named filtered_df.

```
Q1 filtered df['CrimeCount'].quantile(0.25)

Q3-filtered_df['CrimeCount'].quantile(0.75)

IQR=Q3 - Q1

lower bound = Q1 - 1.5 *IQR

upper bound =Q3 + 1.5*IQR
```

outliers filtered_df[(filtered_df['CrimeCount'] < lower bound) | (filtered_df['CrimeCount']> upper_bound)]

print('Outliers:')

print (outliers)

The statement **outliers = filtered_df[(filtered_df['CrimeCount'] < lower_bound) | (filtered_df['CrimeCount'] > upper_bound)]** creates a new DataFrame named outliers that contains only the rows from filtered_df where the 'CrimeCount' column values are considered outliers based on the lower and upper bounds calculated using the Interquartile Range (IQR) method.

mean value filtered_df['CrimeCount'].mean() filtered_df.loc[(filtered_df['CrimeCount'] < lower_bound) | (filtered_df('CrimeCount']> upper_bound), 'CrimeCount'] mean_value

print(filtered_df)

The statement **mean_value = filtered_df['CrimeCount'].mean()** calculates the mean (average) value of the **'CrimeCount' column** in the **DataFrame filtered_df** and assigns it to the variable **mean_value**.

filtered_data **=** filtered_df**.**copy()

scaler **=** MinMaxScaler()

filtered_data**.**loc[:, 'CrimeCount_Normalized'] **=** scaler**.**fit_transform(filtered_data[['CrimeCount']])

print(filtered_data)

The statement **filtered_data = filtered_df.copy()** creates a new DataFrame named filtered_data that is a copy of the original **DataFrame filtered_df.** The **statement scaler = MinMaxScaler()** creates an instance of the MinMaxScaler class from the scikit-learn library. The statement **filtered_data.loc[:,'CrimeCount_Normalized'] = scaler.fit_transform(filtered_data[['CrimeCount']])** normalizes the 'CrimeCount' column in the DataFrame **filtered_data** using the **MinMaxScaler** that was previously instantiated.

selectedFeatured **=** filtered_data**.**drop(['CrimeCount','Year','Day'], axis**=**1)

print(selectedFeatured**.**head())

The statement **selectedFeatures = filtered_data.drop(['CrimeCount', 'Year', 'Day'], axis=1)** creates a new DataFrame named selectedFeatures by dropping the specified columns **('CrimeCount', 'Year', and 'Day')** from the original **DataFrame filtered_data.**

encoded_df **=** pd**.**get_dummies(selectedFeatured, columns**=**['District', 'DutyShift','Month','DayOfWeek'])

print(encoded_df)

The statement **encoded_df = pd.get_dummies(selectedFeatures, columns=['District', 'DutyShift', 'Month', 'DayOfWeek'])** creates a new DataFrame named **encoded_df** by applying one-hot encoding to categorical columns in the DataFrame **selectedFeatures**.

```
# Split the data into features (X) and target (y)

y = encoded_df['CrimeCount_Normalized']

X = encoded_df.drop('CrimeCount_Normalized', axis=1)
```

It are used to define the target variable y and the feature variables X for a machine learning task.

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

type(y_test)
```

The statement **from sklearn.model_selection import train_test_split** imports the **train_test_split** function from the **model_selection** module within the scikit-learn library.

Model 1: Linear Regression Model

```
from sklearn.linear_model import LinearRegression

model = LinearRegression()

model.fit(X_train, y_train)
```

The statement **from sklearn.linear_model import LinearRegression** imports the **LinearRegression** class from the **linear_model** module within the scikit-learn library. This code demonstrates the typical workflow with a linear regression model, where you create an instance of **LinearRegression**, train the model using the **fit** method with training data (**X_train** and **y_train**), and then use the trained model to make predictions on new data (**X_test**).

```
from sklearn.metrics import mean_squared_error

y_pred_lr = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred_lr)

print(f'Mean Squared Error: {mse}')
```

The statement **from sklearn.metrics import mean_squared_error imports the mean_squared_error** function from the metrics module within the scikit-learn library. This is a function within scikit-learn that calculates the mean squared error (MSE), a common metric for assessing the accuracy of regression models. It measures the average squared difference between the predicted and actual values. This code calculates the mean squared error by comparing the actual target values (**y_test**) with the predicted values (**predictions**). The result, stored in the variable **mse**, represents the average squared difference between the predicted and actual values.

Model 2: SVR Model

```
from sklearn.svm import
```

SVR# *Initialize SVR model*

svr **=** SVR()

svr**.**fit(X_train, y_train)

y_pred_SVR **=** svr**.**predict(X_test)

mse **=** mean_squared_error(y_test, y_pred_SVR)

print(f'Mean Squared Error: {mse}')

The statement from sklearn.svm import SVR imports the Support Vector Regression (SVR) class from the Support Vector Machines (SVM) module within the scikit-learn library.

Model 3: MLP Regressor Model

from sklearn.neural_network **import** MLPRegressor

mlp_regressor **=** MLPRegressor(hidden_layer_sizes**=**(100, 50), max_iter**=**1000, random_state**=**42)

mlp_regressor**.**fit(X_train, y_train)

y_pred_MLP **=** mlp_regressor**.**predict(X_test)

mse **=** mean_squared_error(y_test, y_pred_MLP)

print(f'Mean Squared Error: {mse}')

The statement from sklearn.svm import SVR imports the Support Vector Regression (SVR) class from the Support Vector Machines (SVM) module within the scikit-learn library. The statement **svr.fit(X_train, y_train)** is calling the **fit** method of an instance of the Support Vector Regression (SVR) model (**svr**). The statement **y_pred_SVR = svr.predict(X_test)** is using the **predict** method of an instance of the Support Vector Regression (SVR) model (**svr**). Calculates the mean squared error (MSE) between the actual target values (**y_test**) and the predicted values (**y_pred_SVR**) using the **mean_squared_error** function from scikit-learn.

Model 4: Random Forest Regressor

from sklearn.ensemble **import** RandomForestRegressor

rf_regressor **=** RandomForestRegressor(n_estimators**=**100, random_state**=**42)

rf_regressor**.**fit(X_train, y_train)

y_pred_rf **=** rf_regressor**.**predict(X_test)

mse **=** mean_squared_error(y_test, y_pred_rf)

print(f'Mean Squared Error: {mse}')

The statement **from sklearn.ensemble import RandomForestRegressor** imports the **RandomForestRegressor** class from the **ensemble** module within the scikit-learn library. The statement **rf_regressor.fit(X_train, y_train)** is calling the **fit** method of an instance of

the **RandomForestRegressor** class (**rf_regressor**). The statement **y_pred_rf = rf_regressor.predict(X_test)** is using the **predict** method of an instance of the **RandomForestRegressor** class (**rf_regressor**). Calculating the mean squared error (MSE) between the actual target values (**y_test**) and the predicted values (**y_pred_rf**) using the **mean_squared_error** function from scikit-learn.

DISCUSSION

This study purpose is to measure and evaluate the accuracy rate or mean square error of different machine learning models to predict the target variable (crime count) in existence of the feature variables (day, week, month, duty shift, year). In machine learning models, linear regression model, SVR model, MLP Regressor, and Random Forest Regressor model are opted to evaluate the accuracy in predicting the target variable (crime count) in connection with feature variables (day, week, month, duty shift, year). MSE score is calculated which shows the best model in these four models of machine learning algorithm. Before discussing the machine learning models, criminal data of Punjab district is described in terms of mean, standard deviation, and quartile. Thereafter, data preprocessing steps are applied to clean and transform the data of the 15-call system in Punjab which is essential before applying the machine learning models. Four machine learning models in terms of MSE (means square error) are discussed to evaluate the best model which are also supported by past studies.

Table 2.
Descriptive Statistics

	Year	Month	Day	Crime Count
Count	38207	38207	38207	38207
Mean	2022	6.520193	15.727772	40.642291
Std	2022	3.446354	8.805003	38.580422
Min	2022	1	1	1
25%	2022	4	8	15
50%	2022	7	16	27
75%	2022	10	23	55
Max	2022	12	31	371

The statement print (filtered_df.describe()) is used to display summary statistics of a DataFrame using the describe() method in pandas. Table 2 shows that Year, 2022 has average 6 month data (mean=6.520, s.d=3.446) and average of almost 16 days (mean=15.727, s.d=8.80), and number of crime count average that was 40.642 alongwith standard deviation 38.580 which depicts huge deviation of values with mean in comparison of standard deviation of month and days. In 2022 year where minimum one month and one day counts one crime. In 2022 year, maximum 12 months and maximum 31 days count 371 crimes. This is crime information along with year, month and day of 30 districts excluding Lahore district. 25% quartile of the data represents the values below which a given percentage of data fall is 4 month and 8 days that have crime counts 15. 50% quartile of the data represents the values below which a given percentage of data fall is 7 month and 16 days that have crime counts 27. 75% quartile of the data represent the values below which a given percentage of data fall is 10 month, and 23 days that have crime counts 55.

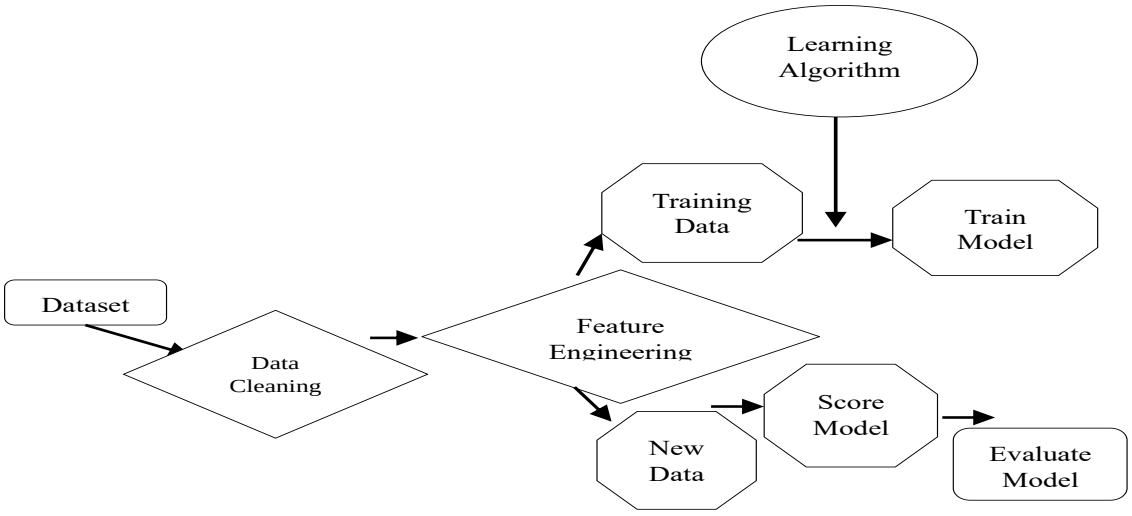

Figure 1.
Data preprocessing steps (George, 2022)

This figure 1 shows the steps of data preprocessing which are followed before applying the machine learning algorithms on the datasets. Data preprocessing steps are followed before applying the models. Outliers are identified by IQR method which are replaced by mean value on the crime data. MinMaxScaler is a feature scaling technique commonly used in machine learning to scale numerical features to a specific range, usually between 0 and 1. This is achieved by computing the minimum and maximum values for each column like crime data and then scaling the values accordingly. MinMax scaling might not be appropriate for all types of data, especially if data has outliers. Therefore, this MinMax scaling is applied after removing the identified outliers in crime data. In pursuing the data preprocessing steps, null value in 8 columns (year, month, day, district, duty shift, crime count, day of week, normalized data of crime count) are identified. No null value is found in these 8 columns.

Following the data preprocessing steps, feature selection is the process of choosing a subset of relevant and significant features from a larger set of features or variables in a dataset. In following the feature selection technique, drop method is applied to create a new Data Frame called selected Featured where specific columns ('CrimeCount', 'Year', and 'Day') from the original DataFrame named filtered_data are removed. Selected features are following: month, district, duty shift, and normalized crime count data.

Following the data preprocessing steps, following data is encoded by one hot encoding technique: district, duty shift, month, and day of week.

Following the data preprocessing steps, data selected after using feature selection technique is used to split into dependent variable (CrimeCount_Normalized) and independent variables (district, duty shift, month, day of week). After splitting the feature variables and target variables, these variables are further divided into training data set and testing data set. Training data is used to build and fine-tune the model, while testing data is used to evaluate its performance on unseen data and estimate how well it will

generalize to new, real-world examples. test_size=0.2 specifies that 20% of the data will be used as the test set, and the remaining 80% will be used as the training set. Following is the training and testing variables of features and target:

X(district)_train: The feature matrix for the training set.

X(district)_test: The feature matrix for the test set.

X(duty shift)_train: The feature matrix for the training set.

X(duty shift)_test: The feature matrix for the test set.

X(month)_train: The feature matrix for the training set.

X(month)_test: The feature matrix for the test set.

X(day of week)_train: The feature matrix for the training set.

X(day of week)_test: The feature matrix for the test set.

Y(CrimeCount_Normalized)_train: The target variable for the training set.

Y(CrimeCount_Normalized)_test: The target variable for the test set.

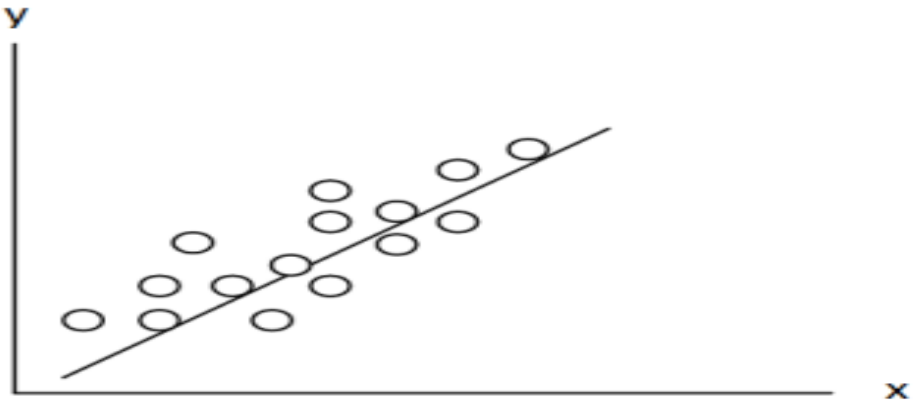

Figure 2.
Linear regression model (Kavitha et al., 2016)

Linear Regression is the most common predictive model to identify the relationship among the variables. Linear regression can be either simple linear or multiple linear regression (Kavitha et al., 2016). Mean Squared Error (MSE) is a common metric used to evaluate the performance of a regression model. Linear regression model shows the MSE score equal to 0.015026. Mean Squared Error (MSE) is a measure of the average squared

difference between the predicted values and the actual values in a regression problem. A smaller MSE indicates that, on average, the model's predictions are closer to the actual values. In contrast, a larger MSE suggests that there are more significant discrepancies between the predicted and actual values. MSE is useful for comparing different models or iterations of the same model. Lower MSE values generally indicate better model performance. Outliers with large errors contribute disproportionately to the overall MSE. The error decreases with increase in sample size as it is normally distributed (Aishwarya, 2022).

Similar to support vector machines, support vector regression (SVR) employs linear kernel functions for regression; however, unlike SVM, SVR sets the tolerance margin (ε) to approximation rather than taking it from the issue (Kavitha et al., 2016). Finding a hyperplane in a high-dimensional space that, while allowing for some error, optimally captures the connection between the input characteristics and the target variable is the fundamental notion underlying support vector regression (SVR). In order to translate the input characteristics into a higher-dimensional space, SVR uses the kernel technique. As a result, complicated connections may be captured by the method without the need for explicit transformation computation. In the high-dimensional space, SVR creates a hyperplane and encircles it with an epsilon-insensitive tube. Deviations from this tube are punished, and data points inside are regarded as well-predicted. The data points that affect the hyperplane's orientation and location are called support vectors. These are the crucial information pieces that either fall inside the margin or go outside of it. The trade-off between decreasing the training error and establishing a smooth fit is managed by the regularization parameter (C). A larger-margin hyperplane, which permits greater deviations inside the epsilon-insensitive tube, is produced by smaller values of C. The width of the epsilon-insensitive tube is represented by epsilon. It establishes the range of mistakes that are acceptable and penalizes departures from it. Two terms make up the loss function in SVR: an epsilon-insensitive loss term that permits a certain degree of inaccuracy, and a regularization term that penalizes complicated models. Moreover, SVMs use the hyperplane that divides the two classes to maximize the margin between them (Pawlak, 1992).

MSE value of SVR is equal to 0.0088280, which is less than MSE score of linear regression model which is equal to 0.015026. This shows that SVR model is best model which is predicting the value of dependent variable (crime count) in the less error in comparison of linear regression model which has higher mean square error while predicting the dependent variable (crime count). SVR is a method that can overcome overfitting. The purpose of overfitting is that data used in the training result in a much better accuracy than the test data. SVR is the development of SVM method with a statistical method that is Regression(Were, 2015). SVR is more superior than the regression method. It can be seen from the results of MSE(Mean Square Error) (Soebroto et al., 2022). The acronym MLP denotes Multi-Layer Perceptron. This is a useful modeling method that employs supervised training with examples of data with known results (Bishop 1995). A neuronal multilayer with a nonlinear activation function makes up a multi-layer perceptron (MLP). Every hidden layer's output may be thought of as a collection of fresh features that are shown to the output layer. Back-propagation is a supervised learning technique used by MLP (Gabralla & Abraham, 2014). One kind of artificial neural network used for regression problems is called an MLPRegressor. A multi-

layer perceptron architecture, a kind of feedforward neural network, is the foundation of the MLPRegressor. An input layer, one or more hidden layers, and an output layer make up its composition. The activation function of every node (neuron) in the MLP adds non-linearity to the model. The sigmoid, hyperbolic tangent (tanh), and rectified linear unit (ReLU) are examples of common activation functions.

To reduce the discrepancy between anticipated and actual values, an optimization approach, usually based on backpropagation, is used to train the MLPRegressor. The network's weights and biases are modified during the optimization phase. A key component of model design is the quantity of neurons in each layer and hidden layers. The model's ability to represent intricate relationships in the data is impacted by these factors. Regression may be carried out using feedforward artificial neural networks, such as multi-layer perceptron. A directed graph including many layers of input nodes (each with an arbitrary number of dimensions) is connected to an output layer (dependent variable) in this paradigm. Backpropagation is used to train the network model. When a neural network has one hidden layer and two outgoing nodes, it generates the lowest mean absolute error among machine learning methods (Yongfei et al., 2022). The outcome is predicted during forward propagation using the activation function and repeated weighted input calculation. In order to determine the gradients, the total error at the output nodes, which was calculated during forward propagation, is transmitted back through the network during backward propagation. In order to lower the error at the output layer, the gradient descent algorithm modifies each weight in the network (Rathore et al., 2018).

The MSE score of the MLP Regressor model is 0.008071, which is lower than the MSE score of the linear regression model (0.015026) and the MSE value of the SVR (0.0088280). This demonstrates that, when compared to linear regression and SVR models, which have larger mean square errors when predicting the dependent variable (crime count), the MLP Regressor model is the best model for predicting the value of the dependent variable with the least amount of error. Backpropagation is a technique used in multilayer perceptron networks to reduce the discrepancy between the desired or targeted output and the received output by iteratively adjusting the weighted and threshold values (Kang, 2017). The efficiency of linear SVM is lower than that of MLP. Their output recall and precision are almost zero, achieving a reasonable level of accuracy. MLP uses a little bit more FPGA resources, but it produces superior accuracy, precision, and recall.

FPGA codes, on the other hand, may further minimize resource consumption by employing optimization techniques, such as threaded and vectorized functions for optimal performance as in Intel's MKL DNN library. This makes FPGA codes very desirable for machine learning applications (Rathore et al., 2018). Unlike application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs) are chips made up of unconfigured logic gates that allow for configuration and reconfiguration capabilities. When the expense of designing and producing an ASIC is prohibitive or when the equipment has to be modified after it is put into operation, FPGAs are used. Electronic instruments, consumer electronics, automobiles, aircraft, and PC equipment specific to a given purpose are just a few of the devices that use FPGAs. FPGAs provide benefits over CPUs, such as precise timing and synchronization, quick decision-making, and concurrent execution of parallel tasks, since they may be used to execute specific computations in hardware (Dase et al., 2006). FPGAs process logic without the need for

an operating system by using floating point accelerators, registers, block RAM, and LookUp Tables (LUTs) (Rathore et al., 2018). Using a variety of distinct decision trees from the training set, the Random Forest Regressor is a prediction-based machine learning technique that outputs the mean prediction of the trees. By combining a large number of decision trees, the random forest (RF) approach reduces variance in comparison to using individual decision trees. It is a type of 'ensemble learning' technique. One of the most potent supervised learning algorithms, random forest can handle both classification and regression problems. An ensemble learning technique called Random Forest regression takes many decision trees and uses the average of their outputs to predict the final result. By generating random subsets of the dataset, Random Forest regression can assist prevent overfitting in the model.

MSE value of random forest regressor model is equal to 0.009667 which is greater than MSE value of MLP Regressor model equal to 0.008071, but less than to MSE score of linear regression model equal to 0.015026 and greater to MSE value of SVR equal to 0.0088280. This shows that MLP Regressor model is best model which is predicting the value of dependent variable (crime count) in the less error in comparison of linear regression model, SVR model, and random forest regressor model which have higher mean square error while predicting the dependent variable (crime count). Random forest regression is also used to try and improve the accuracy over linear regression as random forest will certainly be able to approximate the shape between the targets and features (Ananya et al., 2022). Random forests are ubiquitous among ensemble averaging algorithms because of their ability to reduce overfitting and their efficient implementation (Jason, 2018). Random forest (RF) is an ensemble classification approach that has proved its high accuracy and superiority (Jason, 2018).

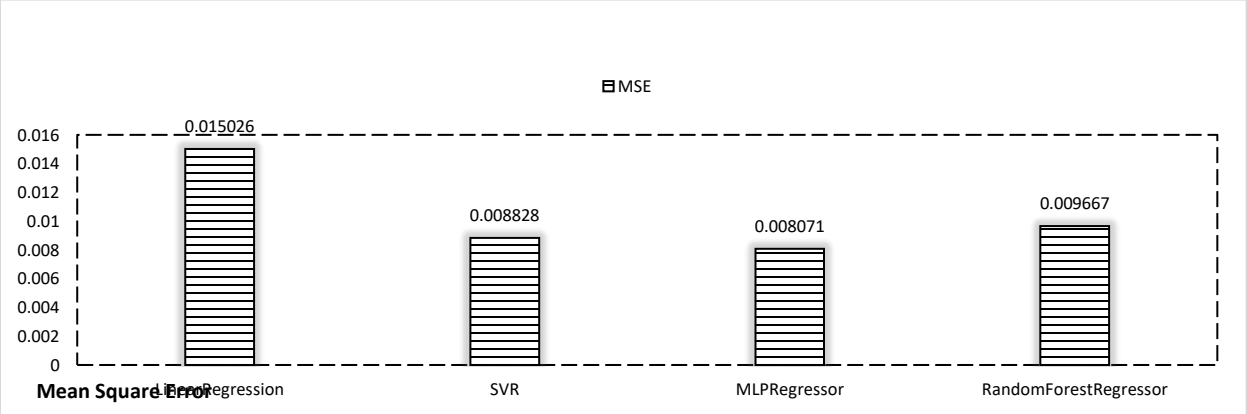

Figure 3.
Mean square error of four regression models

The figure 3 denotes the comparison of mean square error of following regression models: linear regression, SVR, MLP Regressor, and Random Forest Regressor. In order to identify process abnormalities, research examined the effectiveness of a number of supervised machines learning models, including SVM, RF, logistic regression (LR), and radial basis kernel (RBF) classifiers. In order to reduce the computational complexity of learning models, the authors stress the necessity of a feature selection procedure utilizing enhanced rough set theory. They empirically assess this procedure using the application power system dataset (Priyanga et al., 2019).

POLICY IMPLICATION

Crime is defined as any unlawful behavior that lowers the standard of living for individuals. Relevant planning and policy goals, such as the OECD's well-being index, UN Habitat's Safer Cities Program, and the UN Sustainable Development Goals, specifically emphasize the need to create urban places where people feel safe and secure. The Safety Index 2022 depicts that among the safest cities of Pakistan Islamabad comes at 52nd place to be followed by Lahore and Karachi respectively. This also indicates that a lot of work needs to be done to improve the safety and security of the people living in the province. The results of this study will help the police field formations to employ the best machine learning algorithm model to forecast the number of crimes in the Punjab region. The study's conclusions demonstrate that, when predicting the value of the dependent variable (crime count), the MLP Regressor model performs better than the linear regression model, SVR model, and random forest regressor model, all of which have higher mean square errors.

The MLP regressor predicts the number of crimes to occur extremely closely to the actual number of crimes. This MLP regressor also illustrates the significance of characteristics, or independent variables, such as distinct day changes during a week or month in Punjabi regions where the highest concentration of crimes occurs. These characteristics can aid intelligence services in creating a range of plans to stop crimes throughout these changes in the days, weeks, and months in various areas. Selecting the most relevant traits and data is crucial to improving prediction systems' accuracy. The feature selection (FS) strategy, which finds useful characteristics and removes irrelevant ones, often drives the learning algorithm's performance. In addition, FS may save data, storage, and costs while gaining process knowledge. The feature selection (FS) approach is used in conjunction with the data pretreatment stages for the Safe City Lahore crime-related data. The intelligence agency may better comprehend its involvement in reducing crime in Punjab's various districts through feature selection.

LIMITATION

There are drawbacks to some of these machine learning models, despite the fact that they could have several advantages. Since machine learning algorithms need to learn from historical data, they do not always yield precise results or predictions right away. There are restrictions on data availability, resources, technical system performance concerns, and technical data storage. For example, only limited calls (38971) were examined from the total received by Safe City Lahore, and only a small portion of it — specifically, those related to crimes against property — were investigated.

FUTURE DIRECTION

The crime prediction models may also be based on unsupervised learning and on semi-supervised learning. The incorporation of novel features into models for predicting crime is also essential. There needs to be comparison of national crime prediction models in order to identify model commonality. The application of deep learning models, such as transformers, to enhance model performance is also necessary. Future study should be based on creating an optimization model, which will then be used to amass amount of data to produce findings based on a comparison of various machine learning

techniques, including genetic algorithms and deep learning algorithms. Accurate time and gender of caller should be given in addition to crime-prone locations since time is a crucial component of crime analysis and management.

CONCLUSION

The findings show that MLP Regressor model is the best model in predicting the value of dependent variable (crime count) in comparison to other three regression models. Moreover, it has also been observed in crime data of Attock that the crime count was more than 100 on Sunday in the first month during duty shift B, however crime count was less than 20 on the Wednesday in the second month in duty shift C. As per first thousand crime data of Attock, crime count was more than 40 in first ten months of 28 days during duty shift B. However, crime count was less than 10 in Attock in the first ten months of 28 days in duty shift C. These findings assist in controlling the crime rate in the Punjab districts by planning the strategies to combat it in different shifts of days during various months.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the supervisors and their different affiliations.
Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.
Availability of data and material: In the approach, the data sources for the variables are stated.
Authors' contributions: Each author participated equally to the creation of this work.
Conflicts of Interests: The authors declare no conflict of interest.
Consent to Participate: Yes
Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCE

Aishwarya,B. (2022). *Regression Metrics - Of all metrics why MSE?* Retrieved from https://www.linkedin.com/pulse/regression-metrics-all-why-mse-aishwarya-b

Aldossari, B. S., Alqahtani, F. M., Alshahrani, N. S., Alhammam, M. M., Alzamanan, R. M., Aslam, N., & Irfanullah. (2020). *A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. Proceedings of 2020 the 6th International Conference on Computing and Data Engineering.*

Ananya, M., Yash, T.J., Manav, S., & Ramchandra, M. (2022). *Chapter 11 - Impact analysis of COVID-19 news headlines on global economy. Cyber-Physical Systems.* 189-206. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/B9780128245576000017

Antolos, D., Liu, D., Ludu, A., & Vincenzi, D. (2013). Burglary crime analysis using logistic regression. In *Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business, 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part III 15* (pp. 549-558). Springer Berlin Heidelberg.

Aziz, R. M., Sharma, P., & Hussain, A. (2022). Machine learning algorithms for crime prediction under Indian Penal Code. *Annals of Data Science*, 1-32.

Bandekar, S. R., & Vijayalakshmi, C. (2020). Design and analysis of machine learning algorithms for the reduction of crime rates in India. *Procedia Computer Science*, *172*, 122-127.

Biswas, A. A., & Basak, S. (2019). *Forecasting the Trends and Patterns of Crime in Bangladesh using Machine Learning Model. 2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)*.

Cavadas, B., Branco, P., & Pereira, S. (2015). Crime prediction using regression and resources optimization. In *Progress in Artificial Intelligence: 17th Portuguese Conference on Artificial Intelligence, EPIA 2015, Coimbra, Portugal, September 8-11, 2015. Proceedings 17* (pp. 513-524). Springer International Publishing.

Cesario, E., Catlett, C., & Talia, D. (2016, August). Forecasting crimes using autoregressive models. In *2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)* (pp. 795-802). IEEE.

Dase,C., Falcon, J. S., & Maccleery, B. (2006). Motorcycle control prototyping using an FPGA-based embedded control system, *IEEE Control Syst.*, 26, 17-21.

Debanjan, P., & Monisha, C. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons & Fractals. 138.* 109-942.

Elluri, L., Mandalapu, V., & Roy, N. (2019, June). Developing machine learning based predictive models for smart policing. In 2019 IEEE International Conference on Smart Computing (SMARTCOMP) (pp. 198-204). IEEE.

Gabralla, L. A., & Abraham, A. (2014). *Prediction of Oil Prices Using Bagging and Random Subspace. Proceedings of the Fifth International Conference on Innovations in Bio-Inspired Computing and Applications IBICA 2014, 343–354.*

George, L. (2022). *Data preprocessing.* Retrieved from https://www.techtarget.com/searchdatamanagement/definition/data-preprocessing

Jason, M.K. (2018). *Complete analysis of a random forest model.* 1-30. Retrieved from https://www.semanticscholar.org/reader/82ac827885f0941723878aff5df27a3207748983

Javatpoint. (2021). *Regression Analysis in Machine learning.* Retrieved from https://www.javatpoint.com/regression-analysis-in-machine-learning

Kang, N. (2017). "Multi-Layer Neural Networks with Sigmoid Function— "Deep Learning for Rookies (2). Retrieved from https://towardsdatascience.com/multi-layer-neuralnetworks-with-sigmoid-function-deep-learning-for-rookies-2- bf464f09eb7f

Kianmehr, K., & Alhajj, R. (2006). Effective classification by integrating support vector machine and association rule mining. In *Intelligent Data Engineering and Automated Learning–IDEAL 2006: 7th International Conference, Burgos, Spain, September 20-23, 2006. Proceedings 7* (pp. 920-927). Springer Berlin Heidelberg.

Pérez-Enciso, M., & Zingaretti, L.M. (2019). A Guide for Using Deep Learning for Complex Trait Genomic Prediction. *Genes. 10.* 1-19.

Priyanga, S., Gauthama Raman, M. R., Jagtap, S. S., Aswin, N., Kirthivasan, K., & Shankar Sriram, V. S. (2019). An improved rough set theory based feature selection approach for intrusion detection in SCADA systems. *Journal of Intelligent & Fuzzy Systems*, *36*(5), 3993-4003.

Rathore, H., Wenzel,L., Al-Ali, A.L., Mohamed, A., X. Du & Guizani, M. (2018). Multi-Layer Perceptron Model on Chip for Secure Diabetic Treatment. *IEEE Access*, *6,* 44718-44730.

Shohan, F. T., Akash, A. U., Ibrahim, M., & Alam, M. S. (2022). Crime Prediction using Machine Learning with a Novel Crime Dataset. *arXiv preprint arXiv:2211.01551*.

Shraddha, R.B., & Vijayalakshmi, C. (2020). Design and Analysis of Machine Learning Algorithms for the reduction of crime rates in India. *Procedia Computer Science. 172.*122-127.

Soebroto, A. A., Cholissodin, I., Pratiwi, D. E., & P, G. P. W. (2022). Comparative Study of SVR ,Regression and ANN Water Surface Forecasting for Smart Agriculture. *HABITAT, 33*(1), 86–92

Were, K., Bui, D. T., Dick, Ø. B., & Singh, B. R. (2015). A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil

organic carbon stocks across an Afromontane landscape. *Ecological Indicators*, *52*, 394-403.

Yongfei, Q., Chao, Li., Xia, S., & Weigang, W. (2022). *MLP-Based Regression Prediction Model For Compound Bioactivity*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9326362/