* **Methodology:**

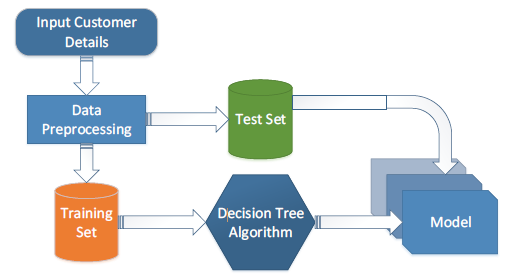
**Dataset:**

Machine learning depends on historical data for understanding past trends, patterns and relationships. In the case of crime analysis, predictions generated by different machine learning models are primarily based on past data related to crimes committed among other factors such as population demographics as well as economic factors. Past data is used in developing predictive models which are more precise compared to others when it comes to machine learning technologies. The utilization of historical data has enabled machine learning algorithms to unveil complex relationships and patterns within data. Through the examination of previous crime trends alongside socio-economic factors and population changes, patterns that have occurred before with potential impacts on future occurrences are identified using algorithms. We gathered historical crime data from the official website of the Bangladesh Police from 2010-2019. Demographics on Bangladesh came from “Worldometer” which also estimated population figures for the same years as mentioned earlier on demographic information above. World Bank database provided us with economic details such as GDP estimates. We also obtained information from the ACLED website, which is a popular group that works with political riots. Once we have put together the datasets, we are given a total of fifteen separate crime datasets which run from 2010 through 2019 where the dataset features are years, population size and GDP, accompanied by the target variable being crime.

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Year | Time interval yearly (int) |
| Population | Total population of a country (int) |
| GDP | GDP per head (int) |
| Crime | Crime rate of a specific crime(int) |

The section serves as a foundation for precise assessment and construction of prediction models in criminology through the emphasis of historical data in machine learning and detailed overview of the corpus assembled.

After creating and setting up the data we split it into two parts so as to gauge how well your models function when analyzing different datasets this way there is some part of information that could still remain fresh. In my thesis I have followed the standard division for any machine learning dataset where 80% is allocated as training dataset while 20% is for testing dataset.



Machine learning models are trained using historical data which is the training set enabling them understand patterns and relationships between input features (for example crime events, population demographics, GDP) and the target variable such as crime rates. The test set is used to evaluate the generalization performance of trained models in isolation.Independent researchers can evaluate whether their models have the potential to make correct predictions from previously unknown data.

**Proposed Model (Regression):**

In machine learning, multiple regression is a statistical technique that is used to establish links between more than one predictor factor and one target variable. This paper on crime analysis in Bangladesh has employed multiple regression in modeling relationships between crime rates and explanatory variables including population, GDP, and year. We selected Multiple regression for several reasons:

1. **Numeric Data:** The dataset consists of numeric variables (population, GDP, year) and aims to predict a numeric target (crime rate), making multiple regression an appropriate choice.
2. **Establishing Relationships:** Multiple regression allows for the examination of how changes in one or more predictor variables (e.g., population, GDP) are associated with changes in the target variable (crime rate). This helps in understanding the influence of socio-economic factors on crime dynamics.
3. **Interpretability:** The coefficients of multiple regression provide insights into the strength and direction of the relationships between predictor variables and the target variable, enabling interpretability of the model.

In multiple regression, the relationship between the predictor variables (X) and the target variable (Y) is represented by the following hyperplane equation:

*Y* = ω0 + ω1.*X1*+ ω2.*X2* + … + ωn.*Xn*

= ω0 +

**Where:**

* *Y* is the target variable (crime rate).
* *X1, X2, ... ,Xn* are the predictor variables (year, population, GDP).
* ω0 is the intercept term (constant).
* ω1, ω2, …, ωn are the coefficients associated with each predictor variable, representing the change in the target variable for a one-unit change in the predictor variable, holding other variables constant.
* **Evaluation Matrix for Regression Model:**
* The necessary metrics and measures have been calculated. These figures include:

**R-square:**

R-squared is a statistical measure that represents the proportion of variance in the target variable (dependent variable) that is explained by the independent variables in the model. It ranges from 0 to 1, where 0 indicates that the model does not explain any variability in the target variable, and 1 indicates that the model perfectly explains all the variability.

*R2* = 1 -

= 1 -

**Mean Squared Error (MSE):**

Mean squared error measures the average squared difference between the actual values of the target variable and the predicted values by the regression model. It provides a quantitative measure of the model's accuracy, with lower values indicating better predictive performance. MSE is calculated by averaging the squared residuals (the differences between actual and predicted values) over the entire dataset.

MSE =

**Mean Absolute Error (MAE):**

Mean absolute error is similar to MSE but calculates the average absolute difference between the actual and predicted values, rather than squaring the differences. MAE provides a measure of the average magnitude of errors in the predictions, regardless of their direction. Like MSE, lower values of MAE indicate better predictive performance, with the advantage of being less sensitive to outliers compared to MSE.

MAE =

=