TABLE OF CONTENT

ABSTRACT

Crime analysis and prediction is a systematic approach for identifying the crime. This system can predict region which have high probability for crime occurrences and visualize crime prone area. Using the concept of data mining we can extract previously unknown, useful information from an unstructured data. The extraction of new information is predicted using the existing datasets. Crimes are treacherous and common social problem faced worldwide. Crimes affect the quality of life, economic growth and reputation of nation. With the aim of securing the society from crimes, there is a need for advanced systems and new approaches for improving the crime analytics for protecting their communities. We propose a system which can analysis, detect, and predict various crime probability in given region. This paper explains various types of criminal analysis and crime prediction using several data mining techniques.

INTRODUCTION

What is Machine Learning?

Machine Learning is a system of computer algorithms that can learn from example through self-improvement without being explicitly coded by a programmer. Machine learning is a part of artificial Intelligence which combines data with statistical tools to pedict an output which can be used to make actionable insights.

The breakthrough comes with the idea that a machine can singularly learn from the data (i.e., example) to produce accurate results. Machine learning is closely related to data mining and Bayesian predictive modeling. The machine receives data as input and uses an algorithm to formulate answers.

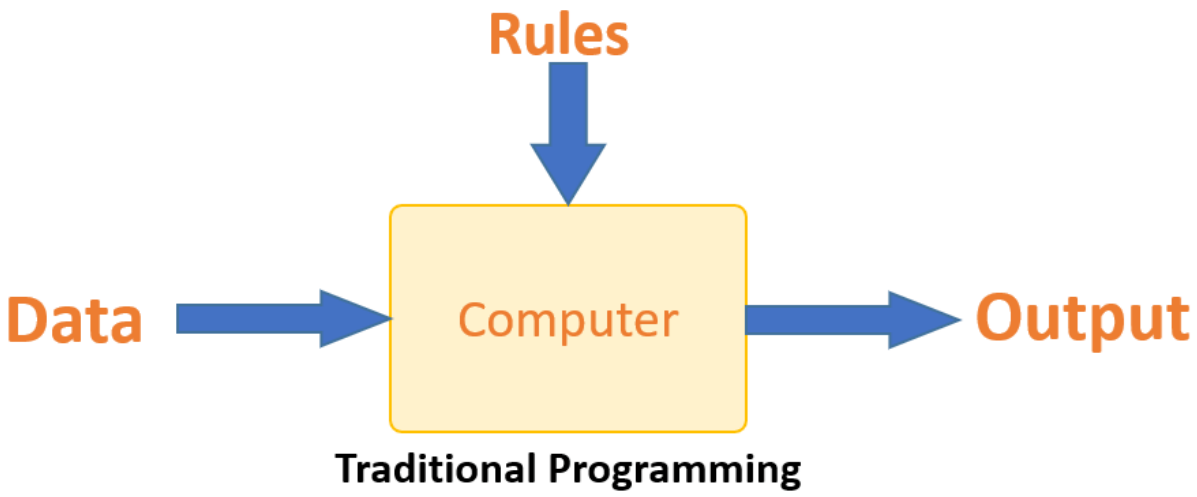
A typical machine learning tasks are to provide a recommendation. For those who have a Netflix account, all recommendations of movies or series are based on the user's historical data. Tech companies are using unsupervised learning to improve the user experience with personalizing recommendation.

Machine learning is also used for a variety of tasks like fraud detection, predictive maintenance, portfolio optimization, automatize task and so on.

Machine Learning vs. Traditional Programming

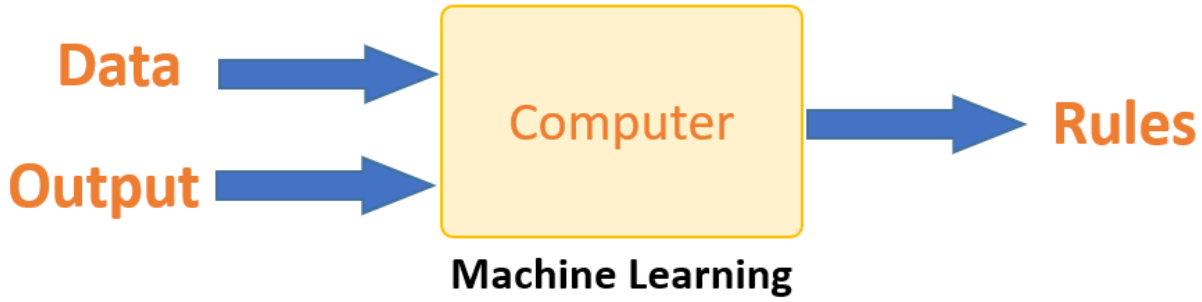
Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.

Traditional programming differs significantly from machine learning. In traditional programming, a programmer code all the rules in consultation with an expert in the industry for which software is being developed. Each rule is based on a logical foundation; the machine will execute an output following the logical statement. When the system grows complex, more rules need to be written. It can quickly become unsustainable to maintain.



Traditional Programming

Machine learning is supposed to overcome this issue. The machine learns how the input and output data are correlated and it writes a rule. The programmers do not need to write new rules each time there is new data. The algorithms adapt in response to new data and experiences to improve efficacy over time.



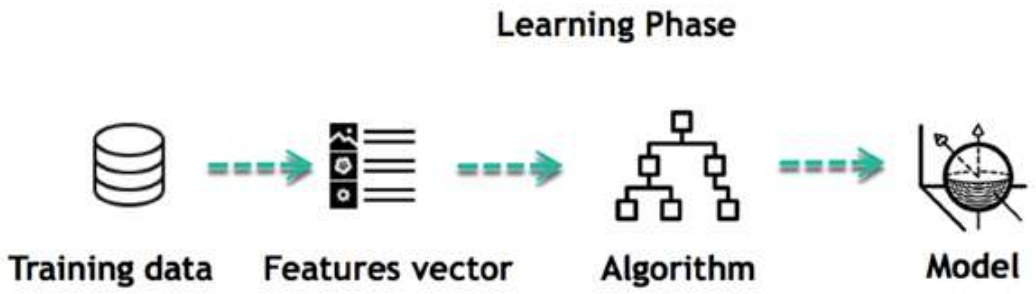
Machine Learning

How does Machine Learning Work?

Machine learning is the brain where all the learning takes place. The way the machine learns is similar to the human being. Humans learn from experience. The more we know, the more easily we can predict. By analogy, when we face an unknown situation, the likelihood of success is lower than the known situation. Machines are trained the same. To make an accurate prediction, the machine sees an example. When we give the machine a similar example, it can figure out the outcome. However, like a human, if its feed a previously unseen example, the machine has difficulties to predict.

The core objective of machine learning is the **learning** and **inference**. First of all, the machine learns through the discovery of patterns. This discovery is made thanks to the **data**. One crucial part of the data scientist is to choose carefully which data to provide to the machine. The list of attributes used to solve a problem is called a **feature vector.** You can think of a feature vector as a subset of data that is used to tackle a problem.

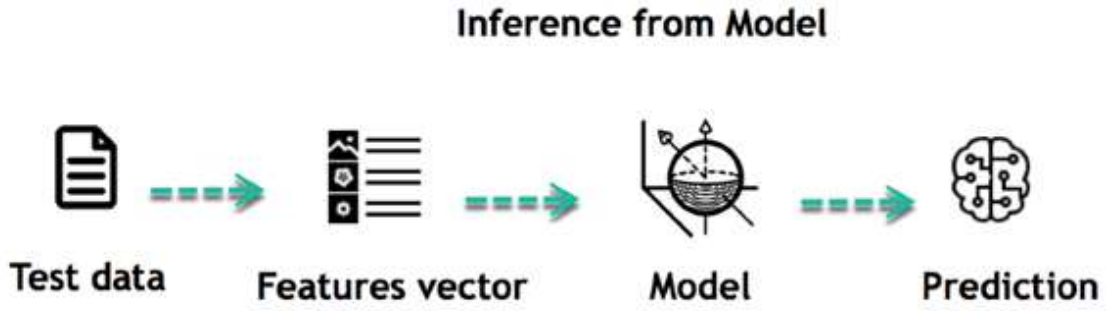
The machine uses some fancy algorithms to simplify the reality and transform this discovery into a **model**. Therefore, the learning stage is used to describe the data and summarize it into a model.



For instance, the machine is trying to understand the relationship between the wage of an individual and the likelihood to go to a fancy restaurant. It turns out the machine finds a positive relationship between wage and going to a high-end restaurant: This is the model

Inferring

When the model is built, it is possible to test how powerful it is on never-seen-before data. The new data are transformed into a features vector, go through the model and give a prediction. This is all the beautiful part of machine learning. There is no need to update the rules or train again the model. You can use the model previously trained to make inference on new data.

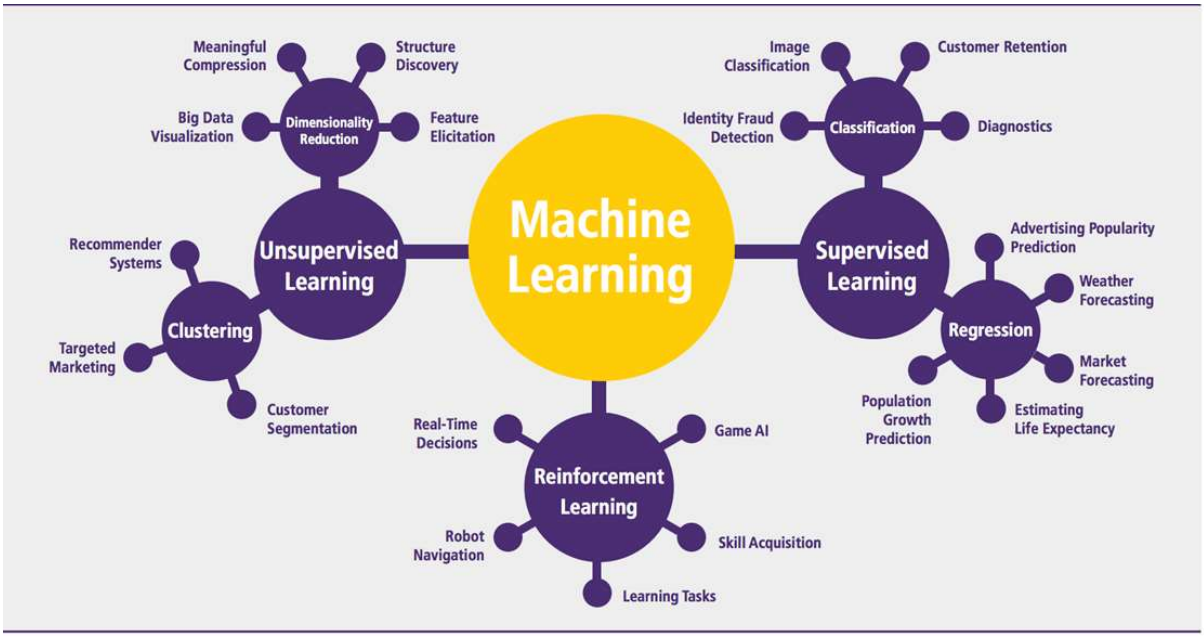


The life of Machine Learning programs is straightforward and can be summarized in the following points:

1. Define a question
2. Collect data
3. Visualize data
4. Train algorithm
5. Test the Algorithm
6. Collect feedback
7. Refine the algorithm
8. Loop 4-7 until the results are satisfying
9. Use the model to make a prediction

Once the algorithm gets good at drawing the right conclusions, it applies that knowledge to new sets of data.

Machine Learning Algorithms and Where they are Used?



Machine learning Algorithms

Machine learning can be grouped into two broad learning tasks: Supervised and Unsupervised. There are many other algorithms

Supervised learning

An algorithm uses training data and feedback from humans to learn the relationship of given inputs to a given output. For instance, a practitioner can use marketing expense and weather forecast as input data to predict the sales of cans.

You can use supervised learning when the output data is known. The algorithm will predict new data.

There are two categories of supervised learning:

- Classification task
- Regression task

Classification

Imagine you want to predict the gender of a customer for a commercial. You will start gathering data on the height, weight, job, salary, purchasing basket, etc. from your customer database. You know the gender of each of your customer, it can only be male or female. The objective of the classifier will be to assign a probability of being a male or a female (i.e., the label) based on the information (i.e., features you have collected). When the model learned how to recognize male or female, you can use new data to make a prediction. For instance, you just got new information from an unknown customer, and you want to know if it is a male or female. If the classifier predicts male = 70%, it means the algorithm is sure at 70% that this

customer is a male, and 30% it is a female.

The label can be of two or more classes. The above Machine learning example has only two classes, but if a classifier needs to predict object, it has dozens of classes (e.g., glass, table, shoes, etc. each object represents a class)

Regression

When the output is a continuous value, the task is a regression. For instance, a financial analyst may need to forecast the value of a stock based on a range of feature like equity, previous stock performances, macroeconomics index. The system will be trained to estimate the price of the stocks with the lowest possible error.

Algorithm Name	Description	Type
Linear regression	Finds a way to correlate each feature to the output to help predict future values.	Regression
Logistic regression	Extension of linear regression that's used for classification tasks. Th output variable 3is binary (e.g., only black or white) rather tha continuous (e.g., an infinite list of potential colors)	Classification
Decision tree	Highly interpretable classification or regression model that splits data feature values into branches at decision nodes (e.g., if a feature is color, each possible color becomes a new branch) until a fina decision output is made	Regression Classification
Naive Bayes	The Bayesian method is a classification method that makes use o the Bayesian theorem. The theorem updates the prior knowledge o an event with the independent probability of each feature that ca affect the event.	Regression Classification
Support vector machine	Support Vector Machine, or SVM, is typically used for th classification task. SVM algorithm finds a hyperplane that optimall divided the classes. It is best used with a non-linear solver.	Regression (not ver common) Classification

Algorithm Name	Description	Type
Random forest	The algorithm is built upon a decision tree to improve the accurac drastically. Random forest generates many times simple decisio trees and uses the 'majority vote' method to decide on which label t return. For the classification task, the final prediction will be the on with the most vote; while for the regression task, the averag prediction of all the trees is the final prediction.	Regression Classification
AdaBoost	Classification or regression technique that uses a multitude of model to come up with a decision but weighs them based on their accurac in predicting the outcome	Regression Classification
Gradient-boosting trees	Gradient-boosting trees is a state-of-the-art classification/regressio technique. It is focusing on the error committed by the previous tree and tries to correct it.	Regression Classification

Unsupervised learning

In unsupervised learning, an algorithm explores input data without being given an explicit output variable (e.g., explores customer demographic data to identify patterns)

You can use it when you do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you

Algorithm	Description	Type
K-means clustering	Puts data into some groups (k) that each contains data with similar characteristics (as determined by the model, not in advance by humans)	Clustering
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	Clustering
Hierarchical clustering	Splits clusters along a hierarchical tree to form a classification system. Can be used for Cluster loyalty-card customer	Clustering
Recommender system	Help to define the relevant data for making a recommendation.	Clustering
PCA/T-SNE	Mostly used to decrease the dimensionality of the data. The algorithms reduce the number of features to 3 or 4 vectors with the highest variances.	Dimension Reduction

AIM AND SCOPE OF THE PRESENT INVESTIGATION

AIM :

OUR AIM TOWARDS THIS PROJECT IS TO PREDICT THE CRIME INCIDENTS THAT HAPPENS IN FUTURE. THE MAJOR ASPECT OF THIS PROJECT IS TO ESTIMATE WHICH TYPE OF CRIME CONTRIBUTES THE MOST ALONG WITH TIME PERIOD AND LOCATION WHERE IT HAS HAPPENED.

SCOPE :

A SYSTEMATIC APPROACH TOWARDS DESCRIPTION AND CLASSIFICATION OF CRIME INCIDENTS

EXPERIMENTAL OR MATERIALS AND METHODS; ALGORITHM USED

MODULES:

- Data Collection
- Dataset
- Data Preparation
- Model Selection
- Analyze and Prediction
- Accuracy on test set
- Saving the Trained Model

MODULES DESCSRIPTION:

Data Collection:

This is the first real step towards the real development of a machine learning model, collecting data. This is a critical step that will cascade in how good the model will be, the more and better data that we get, the better our model will perform.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

Dataset:

The dataset consists of 520 individual data. There are 23 columns in the dataset, which are described below.

1. **ID**: Unique identifier for the record.
2. **Case Number**: The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.
3. **Date**: Date when the incident occurred.
4. **Block**: address where the incident occurred
5. **IUCR**: The Illinois Unifrom Crime Reporting code.
6. **Primary Type**: The primary description of the IUCR code.
7. **Description**: The secondary description of the IUCR code, a subcategory of the primary description.
8. **Location Description**: Description of the location where the incident occurred.
9. **Arrest**: Indicates whether an arrest was made.

10. **Domestic**: Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
11. **Beat**: Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car.
12. **District**: Indicates the police district where the incident occurred.
13. **Ward**: The ward (City Council district) where the incident occurred.
14. **Community Area: Indicates** the community area where the incident occurred. Chicago has 77 community areas.
15. **FBI Code**: Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
16. **X Coordinate**: The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
17. **Y Coordinate**: The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
18. **Year**: Year the incident occurred.
19. **Updated On**: Date and time the record was last updated.
20. **Latitude**: The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
21. **Longitude**: The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.
22. **Location**: The location where the incident occurred in a format that allows for creation of maps and other geographic operations on this data portal. This location is shifted from the actual location for partial redaction but falls on the same block.

Data Preparation:

Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
Randomize data, which erases the effects of the particular order in which we collected and/or otherwise prepared our data
Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
Split into training and evaluation sets

Model Selection:

We used Random Forest Classifier machine learning algorithm , We got a accuracy of 80.7% on test set so