



**T.C.
İSTANBUL UNIVERSITY
INSTITUTE OF GRADUATE STUDIES IN
SCIENCE AND ENGINEERING**



M.Sc. THESIS

SAFE PATH PLANNING WITH MACHINE LEARNING

Yasin Uğur BELEK

Department of Computer Engineering

Computer Engineering Programme

SUPERVISOR

Assist. Prof. Dr. Tolga ENSARI

July, 2018

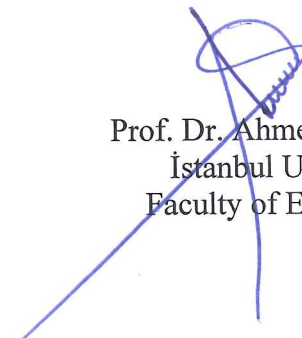
İSTANBUL

This study was accepted on 13/7/2018 as a M. Sc. thesis in Department of Computer Engineering, Computer Engineering Programme by the following Committee.

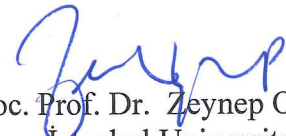
Examining Committee Members



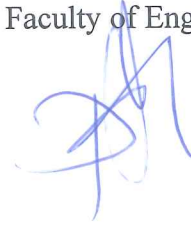
Assist. Prof. Dr. Tolga ENSARI(Supervisor)
İstanbul University
Faculty of Engineering



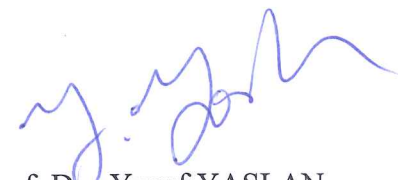
Prof. Dr. Ahmet SERTBAŞ
İstanbul University
Faculty of Engineering



Assoc. Prof. Dr. Zeynep ORMAN
İstanbul University
Faculty of Engineering



Assist. Prof. Dr. Pelin GÖRGEL
İstanbul University
Faculty of Engineering



Assist. Prof. Dr. Yusuf YASLAN
İstanbul Technical University
Computer & Informatics Faculty



As required by the 9/2 and 22/2 articles of the Graduate Education Regulation which was published in the Official Gazette on 20.04.2016, this graduate thesis is reported as in accordance with criteria determined by the Institute of Graduate Studies in Science and Engineering by using the plagiarism software to which İstanbul University is a subscriber.

FOREWORD

This work includes a master thesis study entitled "Safe path planning with machine learning", which is carried out at Istanbul University, Institute of Science and Technology Computer Engineering Department.

During my high school education and preparation of my thesis, my valuable counselor guide me with his knowledge and experience. Assist. Prof. Dr. I would like to express my sincere thanks to Tolga ENSARİ.

Lastly, I would like to thank my beloved mother, father and sisters who have been at every stage of my education life and work during my life and who have never sacrificed their material aid and spiritual support in all circumstances.

July 2018

Yasin Uğur BELEK

TABLE OF CONTENTS

	Page
FOREWORD	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	vii
LIST OF TABLES	viii
LIST OF SYMBOLS AND ABBREVIATIONS	ix
ÖZET	x
SUMMARY	xi
1. INTRODUCTION	1
1.1. GENERAL PARTS	2
1.2. HISTORY OF MACHINE LEARNING	2
1.3. DEFINITION OF MACHINE LEARNING	4
1.4. TAXONOMY OF MACHINE LEARNING ALGORITHMS	5
1.4.1. Supervised Learning	5
1.4.1.1. Regression.....	6
1.4.1.2. Classification	7
1.4.2. Unsupervised Learning	8
1.4.2.1. Clustering.....	9
1.4.2.2. Clustering analysis similarity and distance measures.....	10
1.4.2.3. Dimensionality reduction.....	10
1.4.3. Semi-supervised learning	11
1.4.4. Reinforcement learning	12
1.5. LITERATURE REVIEW	13
1.5.1. Studies for Crime Prediction and Classification.....	14
1.5.2. Studies for Crime Data with Location and Time Predict	14
2. MATERIALS AND METHODS	16
2.1. FEATURES OF DATASET	16
2.2. CRIME DATASET PRE-PROCESSING	17
2.3. FEATURE ENRICHMENT.....	18
2.4. METHODS.....	19
2.4.1. K-Means	19

2.4.2. K-Medians	23
2.4.3. K-Medoids	25
2.4.4. X-Means	26
2.4.5. Expectation-Maximization	29
2.4.6. Distance Measures	30
2.4.6.1. <i>Euclidean Distance</i>	31
2.4.6.2. <i>Manhattan(City-Block) Distance</i>	32
2.4.6.3. <i>Minkowsky Distance</i>	32
2.5. TOOLS	33
2.5.1. Python	33
2.5.2. Google Map API.....	34
2.6. APPLICATION.....	34
3. RESULTS	37
3.1. CHOICE OF K-VALUE	38
3.2. SIMULATION	44
4. DISCUSSION.....	45
5. CONCLUSION AND RECOMMENDATIONS.....	48
5.1. CONCLUSION	48
5.2. RECOMMENDATIONS	48
REFERENCES	50
CURRICULUM VITAE	53

LIST OF FIGURES

	Page
Figure 1.1: Machine learning types	5
Figure 1.2: Supervised learning.....	6
Figure 1.3: Regression algorithms.....	7
Figure 1.4: Classification algorithms	7
Figure 1.5: Unsupervised learning	8
Figure 1.6: Showing four clusters formed from the set of unlabeled data	9
Figure 1.7: Dimensionality reduction.....	11
Figure 1.8: Semi-supervised learning.....	12
Figure 1.9: Semi-supervised learning algorithm	12
Figure 1.10: Reinforcement learning.....	13
Figure 2.1: Fields and descriptions.....	17
Figure 2.2: Data download and pre-processing.....	18
Figure 2.3: K-Means flowchart.....	21
Figure 2.5: Algorithm of K-Medoids	26
Figure 2.6: Expectation-Maximization algorithm	30
Figure 2.8: Euclidean distance.....	31
Figure 2.9: Manhattan distance	32
Figure 2.10: System architecture	33
Figure 2.11: Safe path planning.....	35
Figure 2.12: Example of safe path planning.....	35
Figure 3.1: Application of safe path planning	38

LIST OF TABLES

	Page
Table 2.1: Crime type and value	19
Table 3.1: K-Means	39
Table 3.2: K-Median & Euclidean distance.....	40
Table 3.3: K-Median & Manhattan distance.....	40
Table 3.4: K-Median & Minkowsky distance	41
Table 3.5: K-Medoid & Euclidean distance	41
Table 3.6: K- Medoid & Manhattan distance	42
Table 3.7: K- Medoid & Minkowsky distance	42
Table 3.8: X-Means	43
Table 3.9: Expectation-Maximization	43

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol	Explanation
k	: The number of sets
n	: The number of data points
Q	: Overall value
h	: Non-selected object
i	: Selected object
o_c	: Current medoid
q	: The number of dimensions of the parameters
β_i	: Normalized distance
d	: Distance
p	: The number of variables

Abbreviation	Explanation
AI	: Artificial Intelligence
AIC	: Akaike Information Criterion
ANN	: Artificial Neural Networks
BIC	: Bayesian Information Criterion
CNTK	: Computational Network Toolkit
EBL	: Explanation Based Learning
IUCR	: The Illinois Uniform Crime Reporting code
KDE	: Kernel Density Estimation
KM	: K-Means
K-NN	: k-Nearest Neighbors
ML	: Machine Learning
OLSR	: Ordinary Least Squares Regression
OOP	: Object-oriented Programming
RL	: Reinforcement Learning
SSL	: Semi-supervised Learning
SVM	: Support Vector Machines

ÖZET

YÜKSEK LİSANS TEZİ

MAKİNE ÖĞRENMESİ İLE GÜVENLİ YOL PLANLAMA

Yasin Uğur BELEK

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

Danışman : Dr. Öğr. Üyesi Tolga ENSARİ

Son yıllarda, suç hacmi birçok ülkede ciddi sorun haline gelmiştir. Günümüz dünyasında, suçlular, suç işlemekle ilgili tüm modern teknolojileri ve yüksek teknoloji yöntemlerini en iyi şekilde kullanmaktadırlar. Aynı zamanda, suç miktarı ve modern suçlular daha fazla arttığı için suç verilerinin analiz edilmesi de zorlaşmıştır. Dolayısıyla hükümetler suçları kontrol ve öngörmek için modern teknolojileri ve yöntemleri kullanmaya başlamıştır. Bu yüzden geçmişteki suç kalıplarını inceleyerek hızlı ve verimli bir şekilde insanların yolculuk edebilmesi için iyi bir güvenli yol planlamasına ihtiyaç vardır. Bu çalışmanın asıl amacı güvenli yolu tespit ederek insanların bir yerden başka bir yere giderken daha güvenli bir şekilde gitmesi sağlamaktır. Önceki çalışmalardan farklı olarak, bu tezde suçları analiz ederken gerçek ortamda harita üzerinde makine öğrenimi yöntemleri kullanılarak güvenli yol planlaması yapılmıştır. Bu çalışmada, K-Means, K-Median, K-Medoid, X-Means, Expectation-Maximization yöntemleri gerçek suç verileri üzerinde kullanılmıştır. Deneysel çalışmalarda tezde kullanılan kümeleme teknikleri, birbirleri ile karşılaştırılmıştır. Kullanılan yöntemlerin hangilerinin güvenli yol planlamasında daha iyi sonuçlar verdiği çalışmada gösterilmiştir.

Temmuz 2018, 64 sayfa.

Anahtar kelimeler: güvenli yol planlama, makine öğrenmesi, kümeleme, suç

SUMMARY

M.Sc. THESIS

SAFE PATH PLANNING WITH MACHINE LEARNING

Yasin Uğur BELEK

İstanbul University

Institute of Graduate Studies in Science and Engineering

Department of Computer Engineering

Supervisor : Assist. Prof. Dr. Tolga ENSARI

In recent years, crime rate has become a serious problem in many countries. In today's world, criminals use all modern technology and high-techn methods to commit crimes. In addition, the amount of crime and modern criminals has increased, which makes it difficult to analyze crime data. Thus, governments have enforced the use of modern technologies and methods to control and predict crimes. There is a necessity to plan a good, safe route for people to quickly and efficiently travel by examining past criminal patterns. The basic aim of this study is to determine the safe route and make it safer when people go from one place to another. Unlike previous studies, in this thesis, while analyzing crimes, we perform the safe path planning using machine learning methods on maps in real settings. In this study, the K-Means, K-Median, K-Medoid, X-Means, Expectation-Maximization methods are used on actual crime data. The clustering techniques in the thesis are compared with those in experimental studies. The studies show the methods that yield better results in safe path planning.

July 2018, 64 pages.

Keywords: safe path planning, machine learning, clustering, crime

1. INTRODUCTION

Companies and governments around the world attempt to use artificial intelligence for crime reduction, prevention and response to ongoing crime. The opinions behind most studies are that crimes are compared with predictable. Today, many law enforcement agencies must be able to rank large quantities of crime to find useful patterns. Such data analysis was technologically impossible a few decades ago, but it is possible with the latest developments in machine learning.

Basically, the following problems have been identified: In recent years, increasing crime volume has caused serious security problems in many countries. Law enforcement officers lack the utility software related to the safety of roads. Another important problem is that people should not have any knowledge of whether one is safe when one travels. Finally, there is a lack of information about the safety of living areas.

There are many advantages in the analysis of ML methods and crime data in the real world. For example, law enforcement officers can develop different patrol methods through information that they receive to keep an area safe. With machine learning, one can learn and reveal the way how a crime has occurred. Reports of previous criminal activities and hot crime points can be determined according to time, type or any of them. In the literature, crime analysis, crime prediction, and patrol planning are made for law enforcement officers.

In this study, a different solution is presented. We aim to ensure that passengers travel more safely when they travel to a region. This safe journey can be explained as follows; alternative routes are provided according to the road safety ratios, which correspond to the starting and arrival points determined by the passenger on the map, to determine the more secure route. Safe route planning has been done by analyzing crime data using machine learning according to the region selected by the passenger.

In this study, the data set of Chicago's open verbal data from the city of Illinois was used in the United States. This dataset reflects the reported crime incidents in Chicago since 2001 except for the last seven days [1]. Clustering algorithms from machine learning methods are used. The clustering algorithms are K-Means, K-Median, K-Medoid, X-Means, and Expectation

Maximization. The results of different methods have been compared and the most efficient approach is shown.

This study consists of five sections. In the introduction section, the benefits of machine learning are mentioned in predicting and analyzing the crimes. General information about safe path learning by machine learning is provided.

In the second part of the thesis, the definition and usage of machine learning and methods are explained in detail, and the machine learning models and techniques are examined in a general manner.

The data set is informed in the "Material and Method" section of the third part of thesis.

In the fourth section of the thesis, the data are analyzed, and clustering is performed according to the latitude, longitude, crime type and crime value parameters in the dataset using the machine learning methods, and accuracy amounts are obtained by these methods. The methods are compared in terms of these amounts.

Finally, in the "Discussion and Conclusion" section, the test results are evaluated, the comparison with the studies in the literature is discussed, and future studies are mentioned.

1.1.GENERAL PARTS

Machine learning history, description and application areas are explained. The types of machine learning have been mentioned in detail. Lastly, literature studies related to thesis study have been examined.

1.2.HISTORY OF MACHINE LEARNING

Machine learning can be seen as a new technology, but it is not. In the previous days of artificial intelligence, machine learning study essentially used symbolic data and algorithm design is depending upon consistency. There are many points to start the history of Machine Learning. To begin with, the development process of Alan Turing, the inventor of the computer, from the 1950s to the present day was described [2] [3].

1950 - Alan Turing creates the "Turing Test" to determine whether or not a machine is truly intelligent. In order to pass the test, the machine must be capable of making a human believe that it is another human instead of a computer.

1952 - Arthur Samuel, the pioneer of machine learning, collaborated with IBM to create an ML program that learned to play game Checkers. By the mid-1970's, this program was capable of beating the human players.

1957 - Frank Rosenblatt designed the first neural network for computers (the perceptron), which simulate the thought processes of the human brain.

1967 - "The nearest neighbor algorithm" provides a recognizable pattern of the given data set. Thus, the nearest neighbor algorithm is considered to be a milestone to recognize patterns on a computer.

1979 - The Stanford Cart, the first mobile autonomous robot that allows Stanford students to lift barriers and move at the same time, was invented.

1981 - Explanation Based Learning (EBL) analyses the training information and avoids irrelevant information to form a general rule to follow.

1985 - Neural Network Breakthrough: Researchers (re)discovered the Backpropagation algorithm that allows more complex and powerful neural networks with hidden layers to be trained.

1997 - The computer Deep Blue, by IBM, beats world chess champion Gary Kasparov.

2006 - Computer scientist G. Hinton invented the neural networks' re-branding to explain the "Deep Learning" idiom, the new architectures of deep neural networks that can learn better models.

2011 - IBM Watson computer, developed by IBM, won the Jeopardy TV show in which participants answer their questions in native language.

2012 - The Google team, under the leadership of Andrew Ng and Jeff Dean, has developed a deep neural network that recognizes images and video with Google's infrastructure.

2013 - DeepMind, a British deep learning startup, designed a Deep Reinforcement Learning model to play Atari games that can beat human experts.

2014 - Facebook developed DeepFace DNN, which can recognize people like people.

2015 - Amazon introduced Amazon ML. Microsoft announced the Distributed Machine Learning Toolkit (DMTK), a framework for effectively resolving ML problems on a cluster of GPUs and computers.

2016 - Microsoft releases Computational Network Toolkit (CNTK), its open source deep learning toolkit. Google's AlphaGo program becomes the first Computer Go program to beat an unhandicapped professional human player using a combination of machine learning and tree search techniques.

2017 - Google announces TensorFlow.

1.3.DEFINITION OF MACHINE LEARNING

ML is a science that is a subdivision of artificial intelligence. The aim is to create systems in estimates by making inferences from mathematical and statistical operations and data. In order to solve the problem encountered in everyday life, different methods are applied through intelligent filtering and subjected to complex and logical operations and results are produced. If the problem is not encountered for the first time, it is done with the solution method which is similar to the first. In the learning of the machine, the data sets represent our past cells, while the learning algorithms solve the relations between the data and create the desired probabilistic model and solve the problem by model [4].

“Machine Learning at its most basic is the practice of using algorithms to parse data, learn from it, and then make a determination or prediction about something in the world.” – Nvidia [5].

“Machine learning is the science of getting computers to act without being explicitly programmed.” – Stanford [6].

“Machine learning is based on algorithms that can learn from data without relying on rules-based programming.” - McKinsey & Co [7].

1.4.TAXONOMY OF MACHINE LEARNING ALGORITHMS

This algorithms can be separated into categories according to the method provided for learning the data. Learning categorizations can be divided into four different main categories. Figure 1.1 shows machine learning types [8].

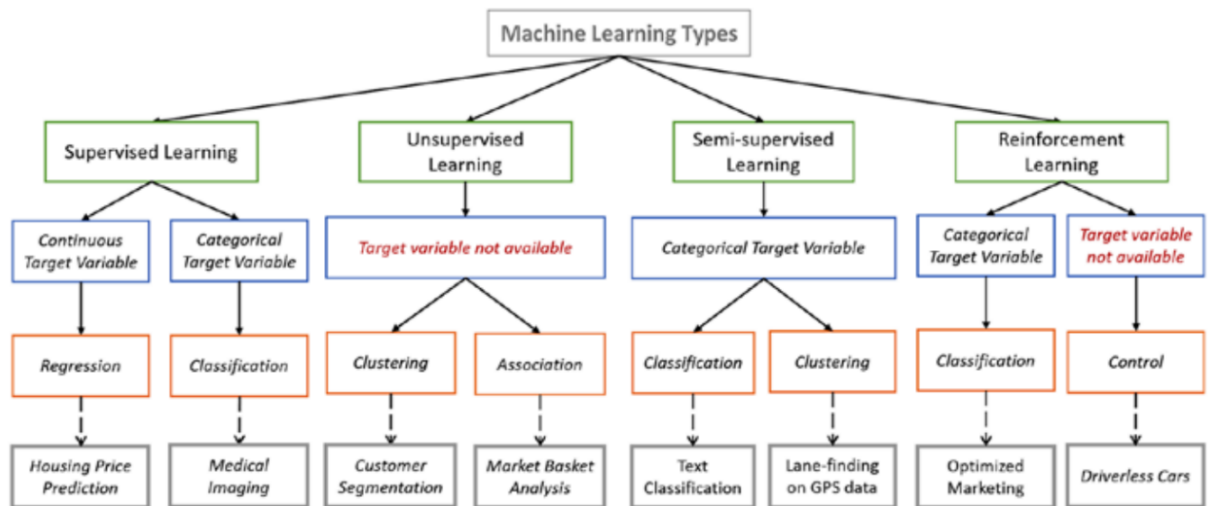


Figure 1.1: Machine learning types

1.4.1. Supervised Learning

Supervised learning substantially ensures a model from the labeled training data, which helps us make predictions on prospective and uncertain data. This approach is similar to the learning under the supervision of a teacher. The teacher provides good examples for the student to memorize, and the student draws general rules from these selected examples.

This algorithm provides a target variable, which is to be predicted from a given set of predictors. Then, using these set of variables, a function that map the entries to the requested outcomes is created. The training process continues until the model reaches an intended degree of accuracy on the training data. Figure 1.2 shows flow of supervised learning [9].

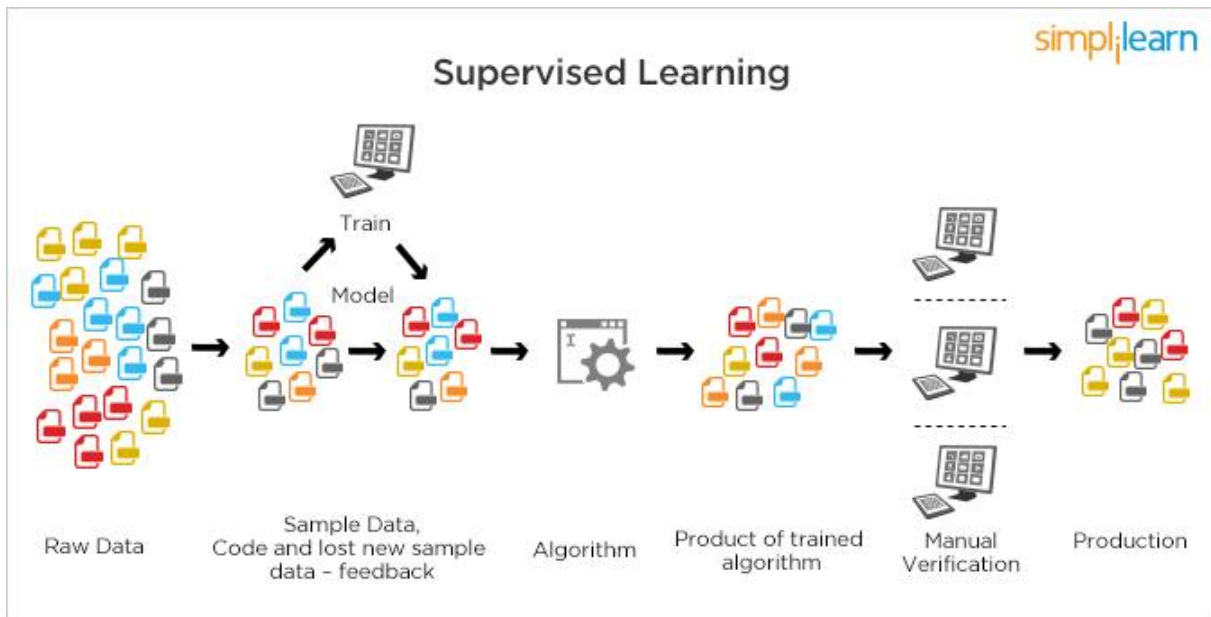


Figure 1.2: Supervised learning

This area has two learning domains: regression and classification.

1.4.1.1. Regression

The machine's ability to recognize numbers, and group them together to form predictions. Supervised learning is the prediction of continuous outcomes, which is also called regression analysis. The main difference between classification and regression models is that the predicted dependent variable has a categorical or continuous value. The field to be estimated is a regression problem if it is a numeric (continuous) variable. A categorical variable is the problem of classification [4]. In regression analysis, many predictive and continuous response variables (outcome or goal) are given, and an attempt is made to find a relationship between these variables that leads to a result.

An example of these variables is the total square meter area of a house, the number of bathrooms it has, and the number of bedrooms it contains. With linear regression, you can estimate the cost of a house by grouping different house samples and learning from their variables and costs. Figure 1.3 shows types of regression algorithm.

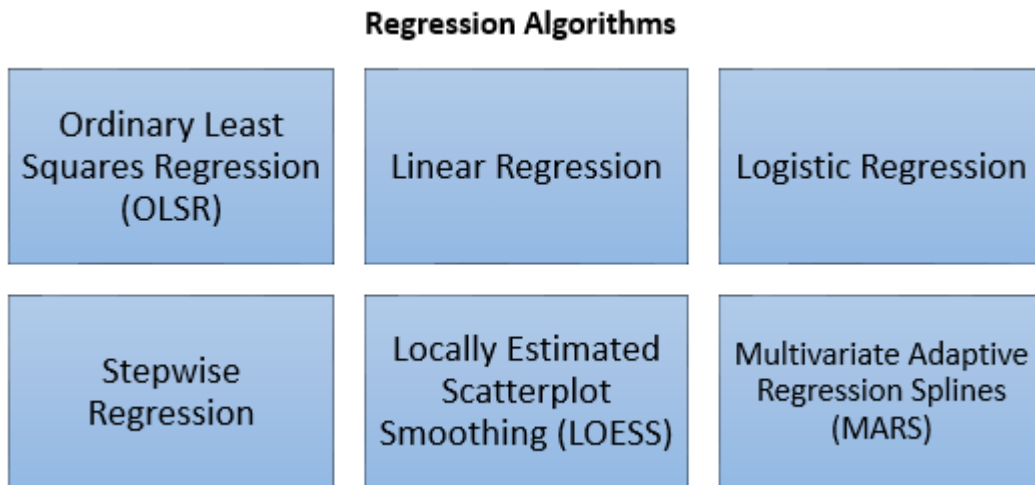


Figure 1.3: Regression algorithms

1.4.1.2. Classification

The machine can identify images or binary things (true and false). The data must be separated into different categories based on the training using past data.

The spam filter is a good example of supervised learning. Our training data includes labels beside emails to indicate whether they are unwanted. The classification is performed according to these labels. The most used algorithm in classification problems is Naive Bayes theorem. Figure 1.4 shows types of classification algorithm.

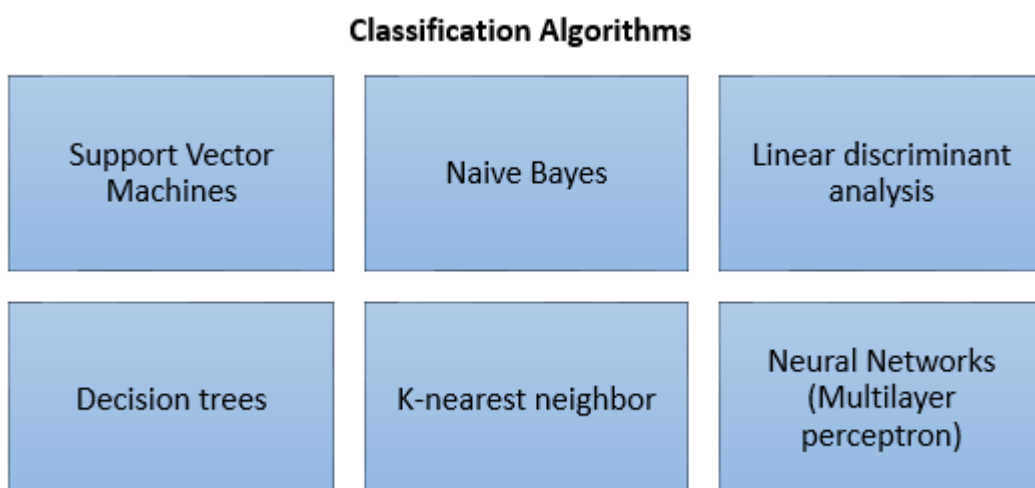


Figure 1.4: Classification algorithms

1.4.2. Unsupervised Learning

This is one of the ML techniques to predict what is unknown in the unmarked data. The class of the input data is unclear. This process extracts the existing but invisible link in the data. The goal is to learn how to do something, and we do not say how to do it using the computer.

Using this algorithm, there is no goal variable to estimate. Thus, the population is clustered in different groups, which is extensively applied to partition customers in diversified groups for specific intervention.

For example, the Turkey population is fragmented into smaller groups with similar demographics and habits of commerce using an advertising platform. With this method, advertisers can achieve their market goals via relevant advertisements. Another example is that Airbnb distinguishes homes when listing houses, which makes it easier for residents to get to their homes while browsing the lists of users. Figure 1.5 shows flow of unsupervised learning [9].

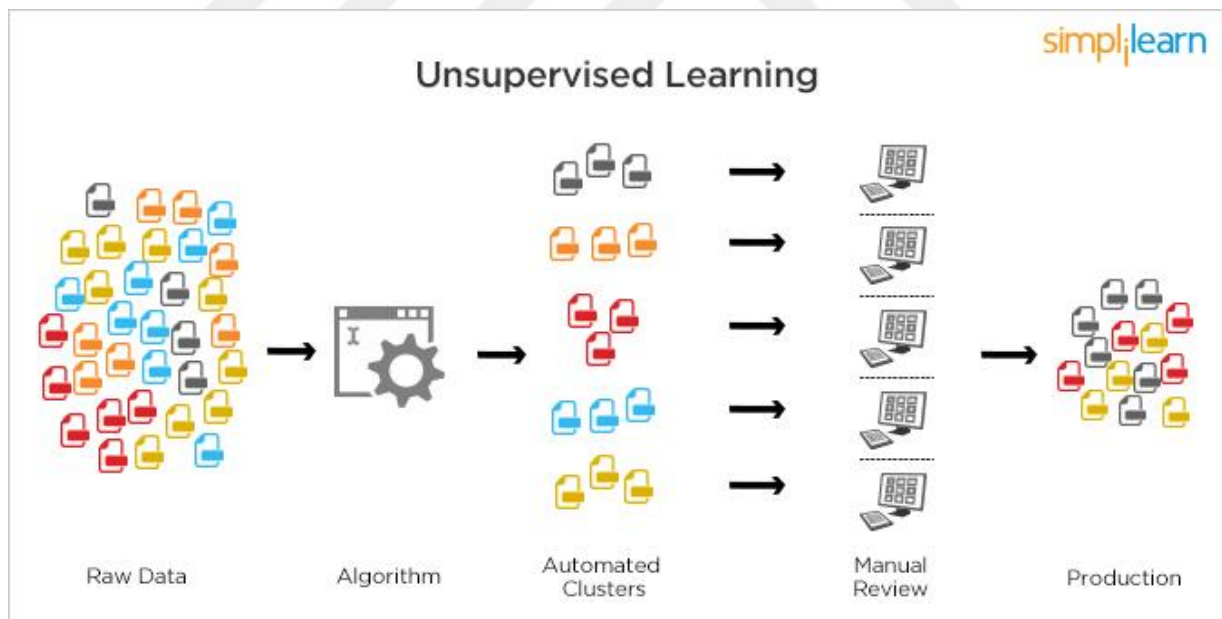


Figure 1.5: Unsupervised learning

Unsupervised learning has been examined in two main categories: clustering and dimensionality reduction.

1.4.2.1. Clustering

Clustering groups are methods to distinguish sub-clusters whose units are not precisely known, but whose variables are almost identical. Through clustering, one can find the appropriate grouping of similar monitors or objects in each cluster. Because clustering does not require the pre-labeling of classes, it is a form of unsupervised learning.

Another definition for clustering is as follows. Clustering enables one to classify the units inspected in a research according to their similarities within certain groups to reveal the common features of the units and make general definitions about these classes (Kaufman and Rousseuw 1990).

Clustering defines the problem class and method class, which is similar to regression. Modeling approaches such as centric and hierarchical types usually regulate clustering methods. All procedures relate to using inherent frameworks in the data to categorize the data with the highest level of relationship and in the optimal manner.

When there is adequate data to constitute clusters and notably when additional data about the members of a cluster can be used to generate further outcomes because of the dependencies in the data, clustering may be beneficial [10]. Some examples of clustering are: news articles are clustered into different types of news; a set of animals is clustered into different types of animals. Figure 1.6 shows clustering [11].

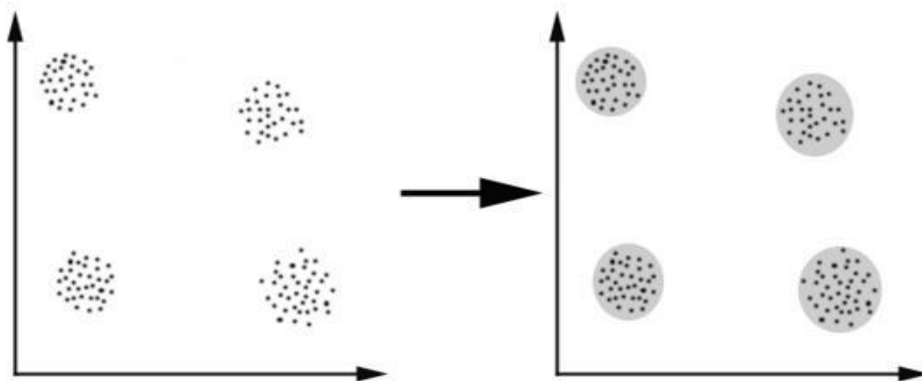


Figure 1.6: Showing four clusters formed from the set of unlabeled data

The most popular clustering algorithms are:

- K-Means
- K-Medians
- K-Medoids
- X-Means
- Expectation Maximization (EM)
- Hierarchical Clustering

1.4.2.2. Clustering analysis similarity and distance measures

There are alternative measures and methods that can be exploited by taking similar distances into account when using the clustering analysis. Using the similarities of these units or their distances to each other pave the way for clustering of units in a dataset. Distance measure or resemblance measure is specified by depending on whether the variables are intermittent or continuous, or whether the variables are at nominal, ordinal, interval, or proportional scale.

Distance Measures

- Euclidean
- Manhattan
- Minkowsky
- Pearson
- Canberra

1.4.2.3. Dimensionality reduction

Another sub-area of SL is dimensionality reduction. It is commonly done with large-size data. In other words, limited storage space and machine learning algorithms can create difficulty for the computational performance of the algorithms. Dimensionality reduction is only as compression. This method simplifies the data while holding as much relevant structure as possible [12]. It can be useful to demonstrate data. For example, a high-dimensional feature set can be presented in 1-, 2- or 3-dimensional feature spaces to visualize using 3D or 2D scatter plots and histograms. Figure 1.7 shows dimensionality reduction [13].

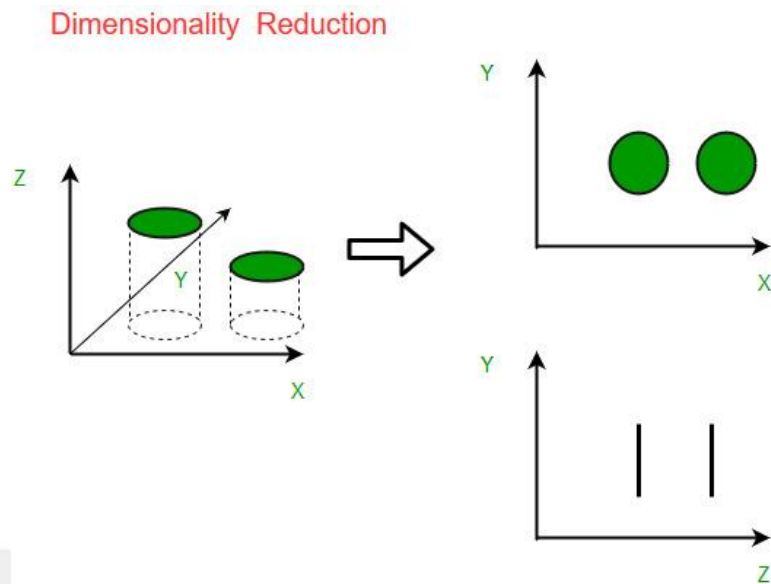


Figure 1.7: Dimensionality reduction

Three common techniques are used: Principal component analysis, Singular value decomposition, Generalized discriminant analysis.

1.4.3. Semi-supervised learning

Which merges both labeled and unlabeled instances to generate a suitable function or classifier [14]. Example problems are classification and regression. The aim is to learn a predictor that forecast next test data better than the predictor learned from the tagged training data only. SSL is supported by its applied value in learning cheaper, quicker, and better. In most of current applications, it is easier to acquire large untagged data x .

Labeled data helps in identification of that there are certain groups of data sorts and what they might be. Then, the algorithm is trained on untagged data to determine the boundaries of those data sorts and may even recognize the new kinds that were unstated in the existing human-inputted tags. For example, Google scans billions of webpages and by processing them, presents classified content to people, properly. Figure 1.8 shows flow of semi-supervised learning. Figure 1.9 shows types of semi-supervised learning algorithm [9].

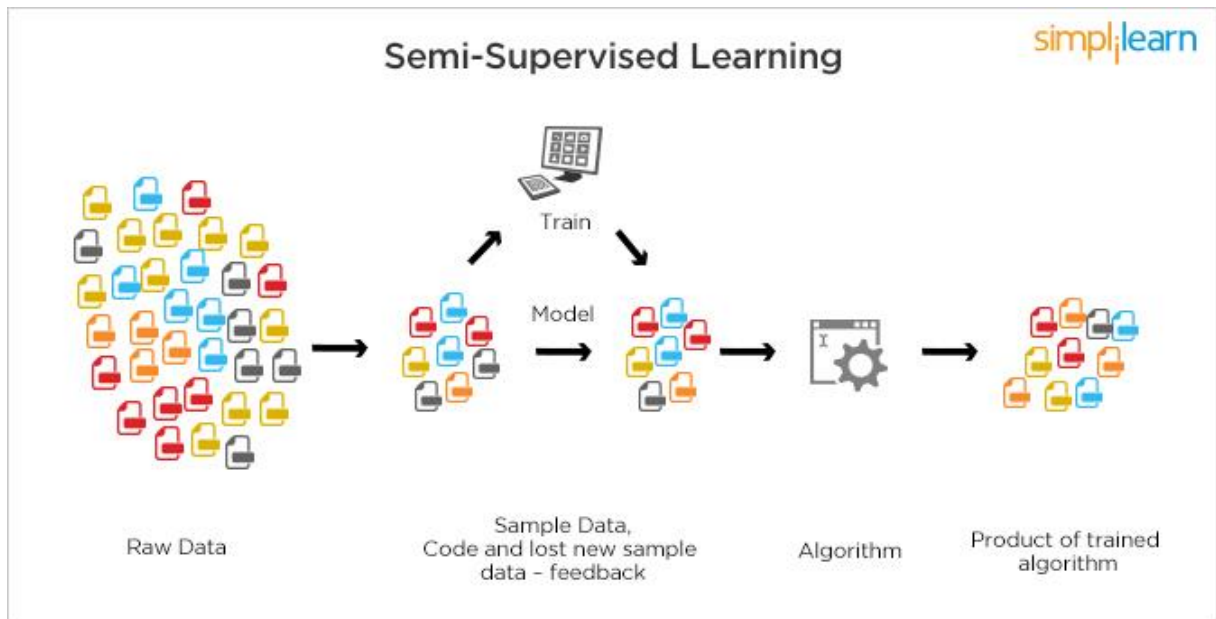


Figure 1.8: Semi-supervised learning

Semi-supervised Learning Algorithm

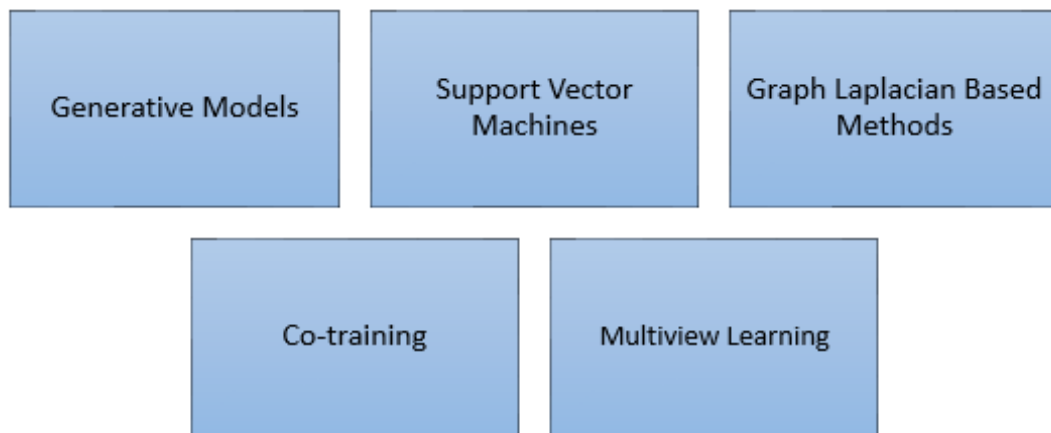


Figure 1.9: Semi-supervised learning algorithm

1.4.4. Reinforcement learning

Reinforcement learning models are the reward-or-punishment way of learning. The RL must decide how to perform the task to perform the reinforcement learning agent when no response key is present. Since training data are not available, it learns from the agent experience. While gathering training data (e.g., this action was good, whereas this action was bad), it attempts to maximize the task in the long run by trial and error. The target is to develop a system that

enhances its performance based on the interactions with the environment. A popular example of reinforcement learning is a chess engine.

The machine is trained to make specific decisions. The machine is exposed to an environment where it continually trains itself using test and error. This machine learns from past experience and tries to capture the best possible information to make accurate business decisions. Figure 1.10 shows flow of reinforcement learning [9].

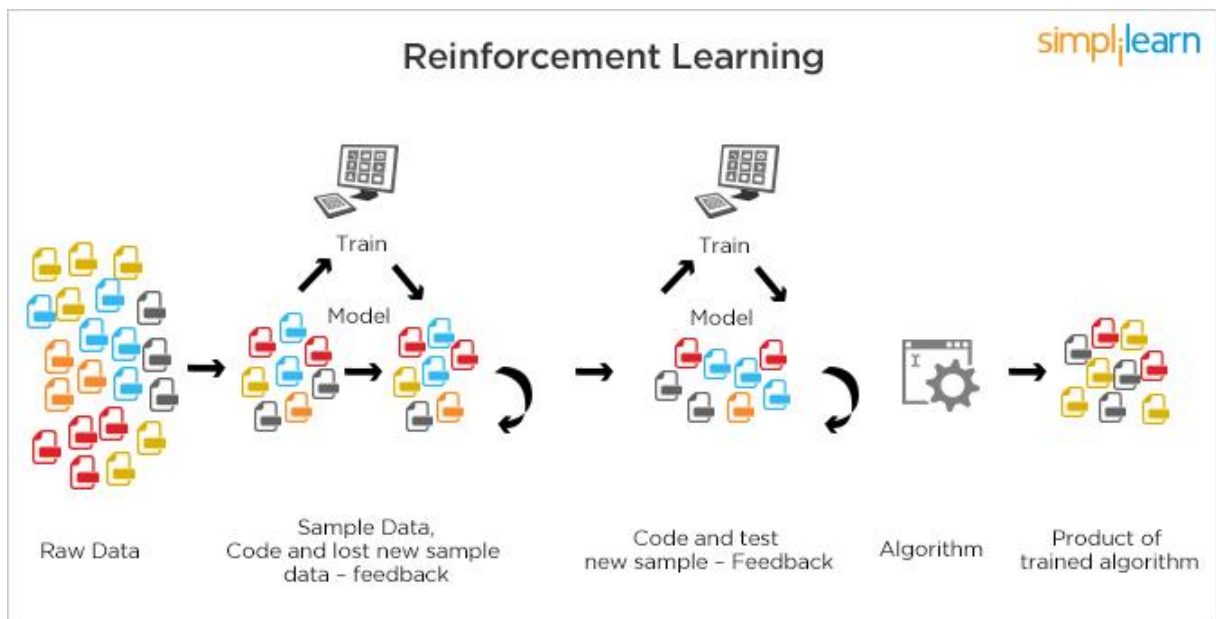


Figure 1.10: Reinforcement learning

Algorithms of Reinforcement Learning

- Q-Learning
- Temporal Difference (TD)
- Deep Adversarial Networks

1.5.LITERATURE REVIEW

Related works to our study can be classified into two groups:

- (i) studies which crime prediction and classification,
- (ii) studies which using crime data with location and time predict.

1.5.1. Studies for Crime Prediction and Classification

The first group examines the crime estimates using different techniques to predict crime. In a recent study by Miquel Vaquero Barnadast [15], supervised machine learning was used to predict crime by examining the crime data of San Francisco. To make fast program executions in a large amount of data, the size of the data can be reduced using K-Means. Then, K-Nearest Neighbors (KNN), CNN, Parzen Windows and Artificial Neural Networks (ANN) algorithms are used for classification. Apparently, the most precise algorithm was ANN to categorize the crime. Although ANN is better than CNN, it does not have better performance.

The subject twitter feelings and crime prediction using weather. This study aims to estimate the time and place for the future of a particular crime scene. The approaches rely on empirical analysis by understanding lexicon-based techniques and classified weather data, which are estimated by the kernel-density estimation and linear modeling based on historical events. As a result, they have proven in the analysis that they can link the crime and predictors of weather and sentiment. For predictive modeling, they used logistic regression. They have improved their prediction performance with the standard hot-spot (KDE) model of theft that they developed [16].

Another study on the empirical investigation of classification algorithms is intrinsic to crime prediction. In this study, a criminal data classification was applied to estimate the "Crime Category" for different states in the U.S. Two distinct classification algorithms were compared to predict different "Crime Categories": Naïve Bayesian and Decision Tree. We have found that Decision Tree performs better than Naïve Bayesian in forecasting all classes of low, medium and high as a result of the study [17].

1.5.2. Studies for Crime Data with Location and Time Predict

In the second group, location and time estimation studies were investigated using crime data. Machine learning has been used to estimate the time and location of crime data. In this study [18], law enforcement forces who have knowledge of criminal patterns at different geological points of a city can take action more efficaciously and react faster. The goal is to use ML techniques to classify a crime event according to the type of event depending on whether it occurs at a certain time or place. San Francisco crimes are investigated. In the classification problem, Decision Tree, Gaussian Naive Bayes, K-NN, Logistic Regression, Adaboost and

Random Forest models are used. As a result of the research, the Adaboost decision tree successfully classifies the criminal activities according to time and place with an accuracy of 81.93%, which is better performance than other machine learning algorithms.

The other study [19] predicts the crime hot-spots using the SVM algorithm. In this study, a Support Vector Machine (SVM) -based approach is proposed to predict the location as an alternative to available modeling perspectives. As a case research in the field of public safety, the focus is on the performance of the one-class Support Vector Machine. The SVM constitutes the next generation of ML methods to find the most appropriate distinction among classes within data sets. The tests on two different spatial data sets (Columbus, St. Louis, USA) show that the one-class Support Vector Machine reveal rational outcomes when we select suitable parameters for the method. The K-Means is useful for data selection. Thus, the Support Vector Machine provides a favorable approach to hotspot crime prediction.

In a recent study by Yu et al. [20], in the United States, a crime estimation model was developed in collaboration with law enforcement officers. Data clusters with place and time information of each crime consist of original criminal registrations and are obtained from the police. In the study, additional spatial and temporal characteristics are extracted from raw data using different classification methods such as Support Vector Machine, J48, and Neural Networks to reveal the method that generates optimal outcomes. The Neural Network estimation predominantly yields better results than other classification methods.

2. MATERIALS AND METHODS

In this section, the thesis work refers to the data set, methods, tools, and finally the application that has been developed.

2.1.FEATURES OF DATASET

In this study, in the United States, an open verbal cluster of data from Chicago, Illinois, was used. The Chicago crime data set is preferred because Turkey has no set of detailed crime data. This data set reflects the reported criminal incidents in the city of Chicago from 2001 to present, except last week. The CLEAR system of the Chicago Police Department is the data provider. To protect the privacy of crime victims, the addresses are only indicated at the block level and not specific locations [1].

There are approximately 6.3 million criminal data points in the Chicago dataset since 2001. Each data point consists of 22 fields as shown in Figure 2.1.

Field	Description
ID	Unique identifier for the record
Case Number	The Chicago Police Department RD Number
Date	Date when the incident occurred in mm/dd/yyyy
Block	The partially redacted address where the incident occurred.
IUCR	The Illinois Uniform Crime Reporting code
Primary Type	The primary description of the IUCR code
Description	The secondary description of the IUCR code, a subcategory of the primary description.
Location Description	Description of the location where the incident occurred.
Arrest	Indicates whether an arrest was made.
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.
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
Beat	Indicates the beat where the incident occurred. A beat is the smallest police geographic area – each beat has a dedicated police beat car. Three to five beats make up a police sector, and three sectors make up a police district. The Chicago Police Department has 22 police districts.
District	Indicates the police district where the incident occurred.
Ward	The ward (City Council district) where the incident occurred.
Community Area	Indicates the community area where the incident occurred. Chicago has 77 community areas.
FBI Code	Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
X Coordinate	The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Y Coordinate	The y coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection. This location is shifted from the actual location for partial redaction but falls on the same block.
Year	Year the incident occurred.
Updated On	Date and time the record was last updated.
Latitude	The latitude of the location where the incident occurred.
Longitude	The longitude of the location where the incident occurred.

Figure 2.1: Fields and descriptions

2.2.CRIME DATASET PRE-PROCESSING

In the knowledge discovery process, the data-preprocessing step, which gives rise to the development of high-quality and powerful data before ML algorithms are applied, has a crucial role. Thus, the data-preprocessing stage is inevitable for the research to obtain clean and

beneficial data. Before implementing machine learning algorithms on our data, we perform a series of preprocessing steps as follows:

Data cleaning: there are several methods for attribute or feature selection. For this study, the manual method was selected for attribute selection, which relied on human knowledge and intellect. The dataset with "Case Number, IUCR, Location Description, Arrest, Domestic, Beat, District, Ward, Community Area, FBI Code, X and Y Coordinate" attributes were removed because these areas were not useful for the current analysis.

Data reduction: There are more than 6 million data points in the crime data set. The amount of data in the dataset was reduced because it is notably difficult to test too much data. The algorithms were only applied to the data in the data set from 2017 to today.

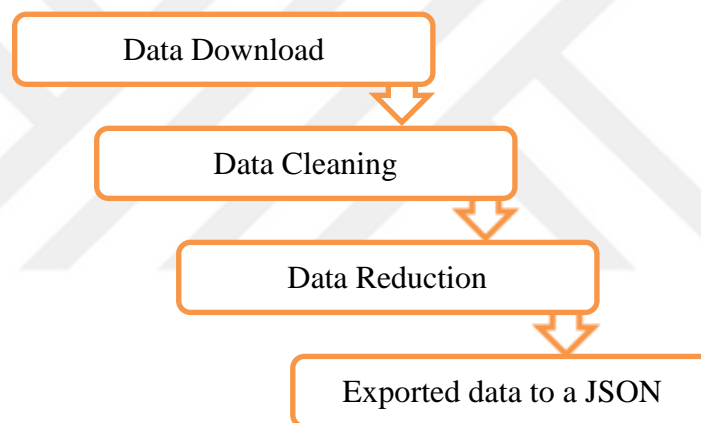


Figure 2.2: Data download and pre-processing

2.3.FEATURE ENRICHMENT

A new feature was added because the information in the dataset for the clustering problem was not sufficient to assess the crime. The dataset was based on the types of crimes that were most committed. A numerical value was assigned according to the specified crime type, and the crime value was named in the attribute. Thus, the use of the crime value (type of number) for use in the type of text calculations is permitted to participate in the calculations. The crime types and values are shown in table 2.1.

Table 2.1: Crime type and value

Crime Type	Crime Value
Public peace violation	1
Intimidation	1
Obscenity	1
Deceptive practice	2
Weapon violation	2
Liquor law violation	2
Offense involving children	2
Gambling	2
Other Offense	3
Criminal Damage	3
Interference with public officer	3
Battery	4
Criminal trespass	4
Assault	4
Robbery	5
Arson	5
Burglary	6
Theft	6
Motor vehicle theft	6
Sex offense	7
Crime sexual assault	7
Narcotics	8
Prostitution	8
Kidnapping	9
Homicide	10

2.4.METHODS

Clustering algorithms from machine learning methods are used. The clustering algorithms used were K-Means, K-Median, K-Medoid, X-Means, Expectation-Maximization and different distance measurement methods.

2.4.1.K-Means

The K-Means algorithm is an elementary unsupervised learning algorithm that solves the famous clustering problem. It is a center-based technique. MacQueen first introduced the K-Means in 1967 [21]. K-Means is the simplest algorithm developed to group a certain number

of kits (k) over a given data set. The letter "k" in the name of the algorithm indicates the number of sets.

The algorithm begins with a random partition of k sets, and the objects are assigned to nearby clusters according to their distance to the cluster centers. Each piece of data can only belong to one cluster. The cluster center is found by taking the arithmetic mean of all coordinates of all points in the cluster (Cui and Potok, 2005). The aim of the algorithm is to minimize the objective function values of the generated clusters.

The steps of the algorithm are as follows:

1. Place K points into the space occupied by the clustered objects. These points denote the preliminary group centroids.
2. Assign each object to the group with the closest centroid.
3. When all objects have been allocated, recalculate the circumstances of the K centroids.
4. Repeat Steps 2 and 3 until the centroids cannot move. This process generates a division of the objects into groups from which the metric to be diminished can be calculated.

In the first step of the algorithm, first, k elements to represent the cluster centers are determined. In the MacQueen algorithm, the cluster centers are selected from the first k elements. However, if the values of the elements are notably close to one another, the selection process is random, or the remote elements are selected. These identified elements are the initial clusters and form the first cluster centers. The cluster member is called the cluster center, whose weighted average value is closest to this value.

In the second step of the algorithm, the specified elements are included in one of the nearest k cluster centers. Various geometric methods are used to find the degree of closeness of the elements to the cluster center. One of them specifies the boundaries among the clusters and determines which clusters are closer to which cluster center. First, two clusters are combined with a central straight line. When another line passes through the midpoint of the merged rectangle and crosses the straight line perpendicularly, this is considered the limit of the right two clusters. By considering the boundary line, it becomes clear the group to include the elements. The most preferred technique to calculate the distance among points is the Euclidean

correlation. In 2-dimensional data, when two cluster centers are correctly used, the plane is used instead of the correct one when the number of dimensions is increased. In multidimensional data, multi-dimensional planes are preferred. In the k-means algorithms developed by computer programs, instead of the planes, the distance among the points is calculated, and the closeness of the points is considered [22].

In the third step, the weighted average of the cluster elements is recalculated with the new additional element in each cluster to form a new cluster center. The weighted average is calculated by taking the average values of all elements of each dimension of the cluster. While the selected elements at the beginning of the algorithm form the cluster center, the new cluster centers in the second loop result are no longer cluster elements and only average values. This new member represents the cluster center in subsequent elections. In each cycle, the elements can be included in a different cluster. Finally, the inclusion of elements into a cluster and the cycle of recalculation of cluster centers continue until the cluster boundaries stop changing. With the KM, a stable set of clusters usually emerges after a few dozen cycles. Figure 2.3 shows K-Means flowchart [23].

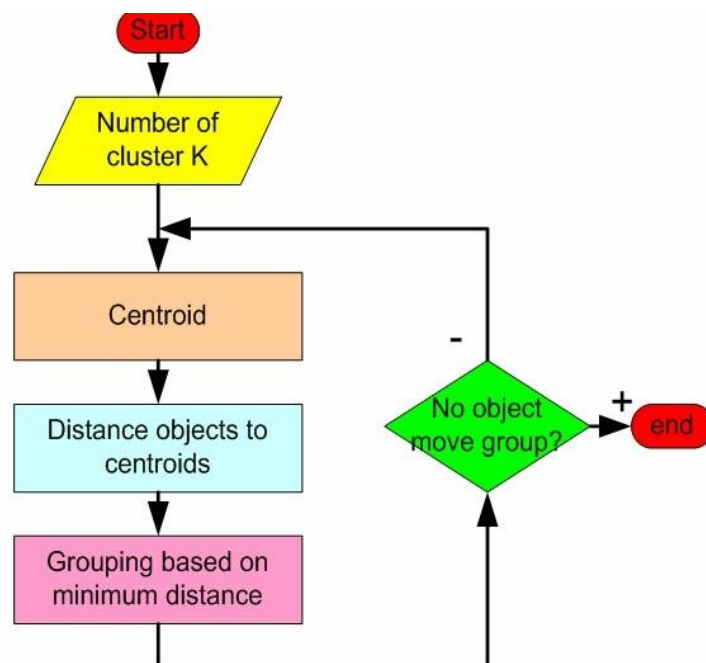


Figure 2.3: K-Means flowchart

The K-Means uses the error squared criterion as the clustering criterion to measure the quality of the clusters formed. This process is accomplished by reducing the squares of the distances among the data point and the associated cluster center. The square of the total error in the K-Means result for the dataset containing N records in K sets is calculated as follows [24].

$$Perf_{KM}(X, C) = \sum_{i=1}^N \min \|x_i - c_j\|^2 \quad j = \{1, \dots, k\} \quad (2.1)$$

In the above equation, x_i is a data point; c_j is the center of j ; k is the number of sets to be formed; n is the number of data points in the data set.

The number k of clusters that are formed according to their closeness to one another is a known constant positive integer before the operation and unchanged by the end of the clustering operation. KM and similar clustering algorithms do not provide a solution to determine the number k . However, in many cases, there is no need to determine a specific k value. Preliminary work has been performed to determine the k value in the analysis phase. Using the estimated value, the clustering algorithm is performed, and the outcomes are evaluated. If the expected clustering does not occur at the end of the evaluation, a re-clustering algorithm may be run using another k value, or the data may be altered.

To calculate the effectiveness of the clusters after each run of the algorithm, the average distance among the records in the cluster is compared to the average distance among the clusters. Different methods can be used in the calculation. However, to determine the final benefit in terms of the application being considered, the clusters must be evaluated on a more subjective basis.

This algorithm is preferred because of the following features

1. The number of clusters is a readable parameter, which makes the analysis flexible.
2. The algorithm is easy to implement and works quickly.
3. Successful results can be obtained in different distributions.
4. Categorical data can be converted into numerical data and brought into a state to work with the KM algorithm.
5. Clustering results can be easily expressed both graphically and in writing.

The K-Means algorithm uses arithmetic and geometric computation techniques to separate the data into sets.

i. K-Means Arithmetic Calculation

In the KM algorithm, there are new cluster centers taking the arithmetic mean of the points in each cluster. The arithmetic mean is obtained by dividing the number of dots by the sum of the dots in a set.

ii. K-Means Geometric Calculation

In this calculation, cluster boundaries are used to divide the data into clusters. In a two-dimensional system, cluster boundaries are correct, while in a three-dimensional system cluster boundaries are plane. In an N-dimensional system, cluster boundaries will be multidimensional planes (hyperplanes).

2.4.2. K-Medians

The K-Means algorithm is identical to the K-Medians algorithm, but the K-Medians are slower and more resistant to contradictions. The K-Medians algorithm is a very important clustering method. The working mechanism of K-Medians is to place n objects in the d -dimensional metric space with the objects in the same cluster and in k clusters more similar than in other clusters.

The steps of the K-Medians algorithm are described below.

```

Q = infinity
do
  for point in dataset
    min = infinity
    index = 0
    for i in k
      dist = distance(point, center[i])
      if dist < min
        min = dist
        index = i
      disjoint-sets.add (index, point)
    for i in k
      center[i] = median(disjoint-set.get(i))
  sum = 0
  for i in k
    for point in disjoint-set.get(i)
      sum = sum + distance (point, center[i])
  oldQ = Q
  Q = sum
while (oldQ - Q) > eps

```

1. Assign each point in the data set to its closest center. The points assigned to the same center are considered in the same cluster, thus they are added to the same disjoint-set. Since each point has been assigned to its closest center, the value of Q will not rise.
2. With the new disjoint-sets as the clusters, calculate their median to determine the updated value of that cluster's center. Because the center is a minimization of 1-norm distances, Q cannot rise as a result of this step.
3. The new value of Q is figured out by collecting all distances between each point and its own cluster center.
4. In the case of the improvements made by this iteration are less than a previously determined epsilon, quit. Otherwise, repeat the process.

Since both steps 1 and 2 can only decrease the overall value of Q , K-Medians is a local optimization algorithm minimizing the sum of the 1-norm distances throughout the dataset [25].

2.4.3. K-Medoids

K-Medoids is based on the finding of k representative objects that represent various structural properties of the K-Medoids algorithm's underlying data. The K-Median is notably similar to the K-Means algorithm. In 1987, Kaufman and Rousseeuw created the K-Medoids algorithm to decrease the vulnerability of the K-Means to noise and exception. The K-Medoids is a clustering algorithm concerned to the K-Means and the medoidshift algorithm. The K-Means and K-Medoids algorithms are divided and attempt to reduce the squared error of the distance between the points tagged to be in a cluster and a point projected as the center of that cluster. K-Medoids selects data points as centers, in opposition to the K-Means algorithm. The difference between the mean and the median is similar to the difference between K-Means and K-Medoids: the mean is the average value of all data items gathered, whereas the median is the value around which all data items are evenly scattered.

The fundamental strategy of K-Medoids is to reveal k clusters in n objects by first irrelevantly finding a representative object (the medoid) for each cluster. Each residual object is clustered with the medoid to which it is the most identical. The technique uses representative objects as datum points instead of referring the mean value of the objects in each cluster. The algorithm takes input parameter k as the amount of clusters to be divided among a set of n objects. A medoid of a limited dataset is a data point from this set, whose approximate dissemblance to all data points is minimal, i.e., it is the most centrally positioned point in the set.

The K-Medoids is productive for a small dataset but does not correctly scale for a huge dataset. Among the divided methods, K-Medoids is an algorithm that produces better and stable clustering results. There is no effect on the order of processing of the data and the clusters of the first assignments. The superiority of K-Medoids compared to K-means is the lowest value of dissimilarity. The processing steps of the K-Medoids algorithm are as follows:

The algorithm happens in two stages:

- BUILD-stage: Using the starting medoids, sequentially k selects "centrally located" objects.
- SWAP-stage: When the object function can be reduced by replacing a selected object with an unselected object, swapping is performed in this case. This continues until the objective function can not be reduced. The algorithm is as bellows:

1. Initially select k random points as the medoids from the given n data points of the data set.
2. Correlate each data point to the nearest medoid by using any of the most common distance metrics.
3. Compute the sum swapping cost TC_{ih} , for each couple of non-selected object h and selected object i . If $TC_{ih} < 0$, i is changed with h .
4. Iterate the steps 2-3 till there is no change of the medoids.

There are 4 circumstances to be considered in this operation:

- (i) Shift-out membership: an object p_i may necessarily be changed from presently regarded cluster of o_j to the other cluster;
- (ii) Update the current medoid: a new medoid o_c is established to substitute for the current medoid o_j ;
- (iii) No change: objects in result of the current cluster possess the unchanged or even smaller square error criterion measure for all potential reallocations regarded;
- (iv) Shift-in membership: an external object p_i is appointed to the current cluster with the substituted medoid o_c . Figure 2.5 shows algorithm of K- Medoids [26].

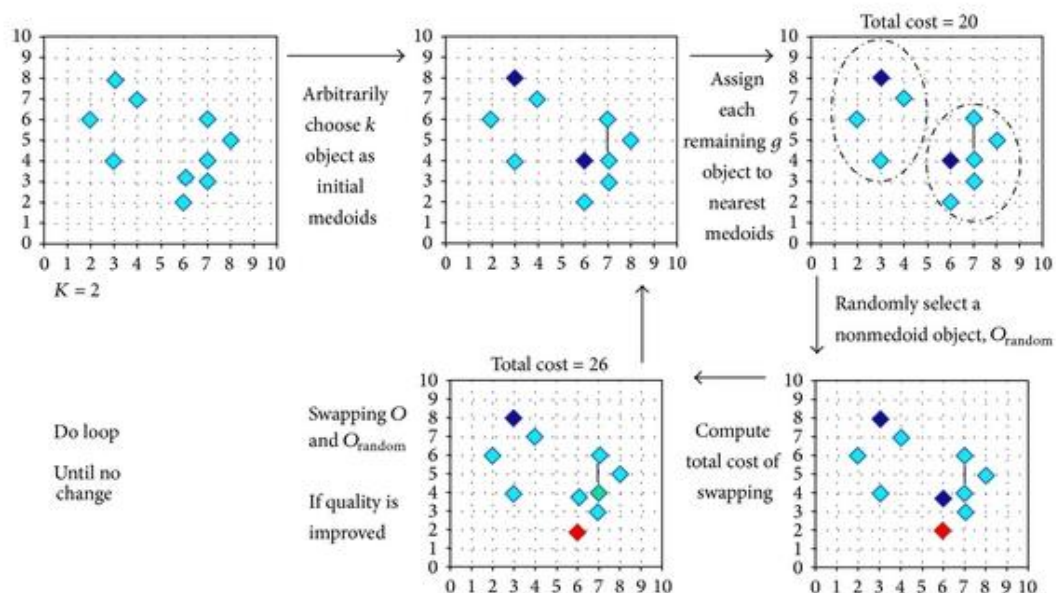


Figure 2.4: Algorithm of K- Medoids

2.4.4. X-Means

X-Means is an alteration of K-Means, which purifies the cluster assignments by recurrently attempting subdivision and holding the optimific separations, until a criterion such as the

Bayesian information criterion (BIC) or Akaike information criterion (AIC) is attained in statistics and data mining [27]. The goal is to determine intrinsic grouping in a set of untagged data. This method provides a fast and efficient way to cluster unstructured data, the use of concurrency speeds up the model construction process, and the use of the Bayesian Information Criterion provides a mathematically sound measure of quality.

In 1978, Gideon E. Schwarz developed the Bayesian information criterion [28]. The BIC or Schwarz criterion is a criterion for model preference in a limited set of models; the model with the lowest BIC one is selected. It is partly grounded on the likelihood function and closely related to Akaike information criterion.

The processing stages of the X-Means algorithm [29] are shown below:

1. Fix the first number of clusters to be k_0 , which should be fairly small.
2. K-means algorithm is applied to all data, while setting $k = k_0$. The separate clusters are appointed C_1, C_2, \dots, C_{k_0} .
3. Iterate the procedure from Stage Four to Stage Nine by exercising $i = 1, 2, \dots, k_0$.
4. For a cluster C_i , apply k-means algorithm by fixing $k = 2$. The separate clusters are appointed $C_i^{(1)}, C_i^{(2)}$.
5. The following p -dimensional normal distribution for the data x_i included in C_i is assumed:

$$f(\theta_i; x) = (2\pi)^{-\frac{p}{2}} |V_i|^{-\frac{1}{2}} \exp \left[-\frac{(x - \mu_i)^t V_i^{-1} (x - \mu_i)}{2} \right] \quad (2.2)$$

And then compute the BIC as

$$BIC = -2 \log L(\hat{\theta}_i; x \in C_i) + q \log n_i \quad (2.3)$$

where $\hat{\theta}_i = [\hat{\mu}_i, \hat{V}_i]$ is the maximum likelihood predict of the p -dimensional normal distribution, μ_i is p -dimensional means vector, and V_i is $p \times p$ dimensional variance-covariance matrix. Moreover, q is the number of dimensions of the parameters and comes in $2p$ if we theorize that the covariance components of V_i are zeros. In conclusion, n_i is the number of elements compromised in C_i and $L(\cdot)$ is the likelihood function.

6. P -dimensional normal distributions with parameters is theorized as θ_i^1 and θ_i^2 for C_i^1 and C_i^2 , severally. The probability density function of this 2-divison model becomes

$$g(\theta_i^1, \theta_i^2; x) = \alpha_i [f(\theta_i^1; x)]^{\delta_i} [f(\theta_i^2; x)]^{1-\delta_i}, \quad (2.4)$$

where α_i is a fixed that is approximated hereinbelow:

$$\alpha_i = \frac{0.5}{K(\beta_i)} \quad (2.5)$$

where β_i is a normalized distance between the two clusters and is derived by

$$\beta_i = \sqrt{\frac{\|\mu_1 - \mu_2\|^2}{|V_1| + |V_2|}}, \quad (2.6)$$

and $K(\cdot)$ shows the lower potential of a normal distribution. The Bayesian information criterion for the model is

$$BIC^l = -2 \log L^l(\hat{\theta}_i^l; x \in C_i) + q^l \log n_i, \quad (2.7)$$

where q^l is obtained by $2 \times 2p = 4p$.

7. If $BIC > BIC^l$, we choose the 2-division model and establish to go on the division. We set $C_i \leftarrow C_i^l$. As the operation for C_i^l , we push the p -dimensional data, the cluster centers, the log likelihood, and the BIC onto the stack and return to Stage Four.
8. If $BIC \leq BIC^l$, we do not choose further clusters and stop. The stacked data that is stored in Stage Seven is extracted and fix $C_i \leftarrow C_i^l$. Then Stage Four is returned. When the stack is empty, return to Stage Nine.
9. The 2-division operation for C_i is completed. We renumber the cluster identification such that it would be without equal in C_i .
10. The 2-division operation for first k_0 divided clusters is finalised. We renumber the all clusters' identifications such that the identifications become without equal.
11. Outcome the cluster identification number to which each element is distributed, the number of elements in each cluster and the each cluster's center.

2.4.5. Expectation-Maximization

In 1977, A. Dempster, N. Laird and D. Rubin expressed an algorithm called the Expectation Maximization (EM) algorithm [30]. The Expectation Maximization is a paradigm of unsupervised, semi-supervised, or lightly SL. The EM is a center-based clustering algorithm, which is a type of frequently used partitioning algorithms. The EM algorithm is used in statistics to find the maximum probability prediction of the parameters, particularly in the probability model. The EM algorithm is used to predict parameters in the presence of unknown data. The EM algorithm and its kinds are the most commonly used algorithms to predict the parameters of Gaussian mixture models. It is a fast-running algorithm, but the results may not have high consistency. In addition, the Expectation-Maximization algorithm is similar to the K-Means algorithm in many respects and can even be considered a K-Means extension.

The fundamental opinion behind the EM algorithm is to use an upper bound function on the negative log-likelihoods of the observed variables by introducing distributions over the hidden variables. This bound is a function of the negative log-likelihoods of the joint distributions of both hidden and observed variables and the introduced distributions over the hidden variables. Fundamental pattern of the Expectation-Maximization:

1. Init-step: random values are allocated to parameters of the model.
2. In E-step, assign points to the model in most compatible way.
3. Update the parameters the allocated the points, previously in M-step.
4. Iterate until parameter values converge

The Expectation-Maximization comes out of two steps called the E-step as well as the M-step.

The Expectation-Maximization loop rotates between performing an expectation (E) step, which forms a function for the expectation of the log-likelihood commented using the current prediction for the parameters, and a maximization (M) step, which calculates parameters maximizing the expected log-likelihood found on the E step. Then these parameter-estimates are used to detect the allocation of the hidden variables in the following E step.

The Expectation-Maximization is frequently utilized for data clustering in ML and computer vision. In natural language processing, two profound examples of the algorithm are the Baum-Welch algorithm for latent Markov models, and the inside-outside algorithm for unsupervised

initiation of probabilistic context-free grammars. Figure 2.6 shows algorithm of Expectation-Maximization [31].

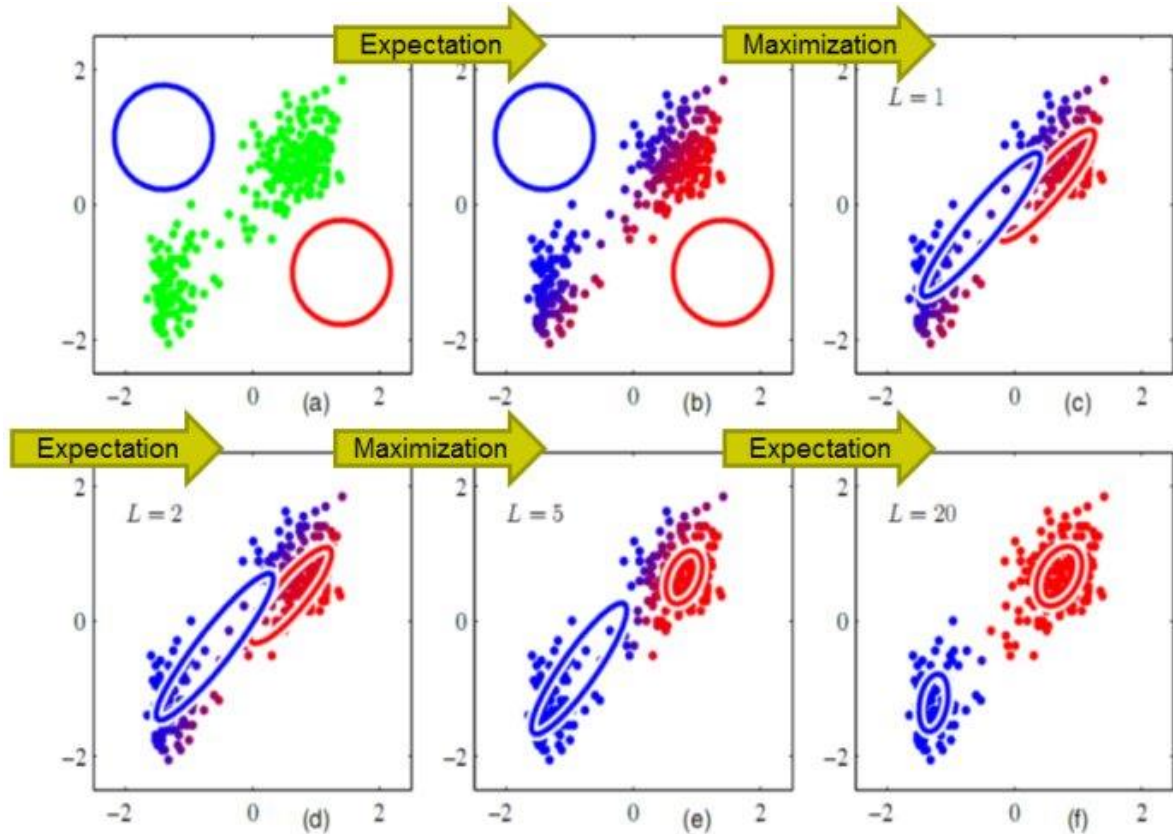


Figure 2.5: Expectation-Maximization algorithm

2.4.6. Distance Measures

The grouping of similar objects is clustering. A type of measure that specifies whether two objects are similar is required. There are two types of measures: distance and similarity measures. The distance measure is preferred by many to examine any pair of objects. When we want to make a cluster analysis on a dataset, different results can appear using different distances, so it is notably important to be careful in selecting the distance.

Distance measure or similarity measure is specified by depending on if they are at nominal, ordinal, interval, or proportional scale.

We can list the general properties of the distance function as: Distance Function's Characteristics;

$d(i, j) \geq 0$; The distance is not negative.

$d(i, i) = 0$; Each unit has zero distance to itself.

$d(i, j) = d(j, i)$; The distance function is symmetric.

$d(i, j) \leq d(i, h) + d(h, j)$; the distance between two units can not be less than the sum of the distances of these two units from a third one [32].

In this study, Euclidean, Manhattan, Minkowsky methods were used for distance between K-Median and K-Medoids.

2.4.6.1. Euclidean Distance

This formulas are, however not standardized data is calculated. This distances are not influenced by the plusage of new objects that may be unsung in clustering analysis. But, differences in scale between dimensions profoundly influence Euclidean distances. The most frequently used distance formulae is the Euclidean one [33].

The distance between two units, n is the number of units and p is the number of variables; $i, j = 1, 2, 3, \dots, n$, i . ve j . the distance of the unit to each other is computed by hereinbelow formula in exercising the Euclidean distance measure. Figure 2.8 shows Euclidean distances [34].

$$d(i, j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \dots + (x_{ip} - x_{jp})^2} \quad (2.8)$$

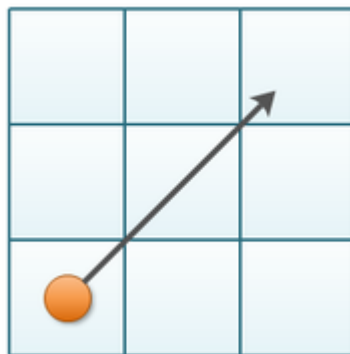


Figure 2.6: Euclidean distance

2.4.6.2. Manhattan (City-Block) Distance

It may appear in the literature with names such as the city block distance or absolute value distance. City block distances track a route along the non-hypotenuse sides of a triangle. The noun signals to the grid-like layout of most American cities almost preclude one from directly going between two points. The Manhattan distance is a metric, where the distance between two points is the total of the absolute distinctions of their Cartesian coordinates. In short, it is the overall amount of the difference between x-and y-coordinates. The Euclidean and squared Euclidean metrics are more affected by outliers than this metric.

When there is a interrelation between the variables in the study, clustering grounded on the distance measurements computed by Manhattan distance would not be purposeful. In the other options, if the units of the measured variables are not same, purposeful outcomes could not be taken from the Manhattan in comparison with the normalized black Euclidean distance [35]. Figure 2.8 shows Manhattan distances [34].

$$d(i, j) = \sum_{k=1}^n |X_{ik} - X_{jk}| \quad (2.9)$$

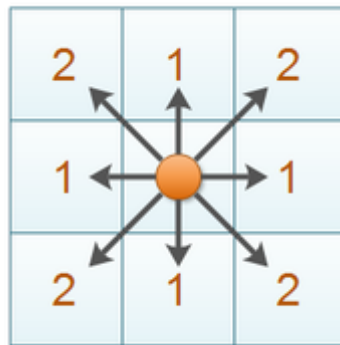


Figure 2.7: Manhattan distance

2.4.6.3. Minkowsky Distance

The distance of Minkowski is a common form. Nevertheless, new formulas are stemmed from different values that p get in the formula. Using the Minkowski distance measure, the distance between two units (x, y) is calculated by the formula.

$$d(i, j) = \sum_{k=1}^n |X_{ik} - X_{jk}| \quad (2.10)$$

When $p = 2$, it would be the Euclidean and when $p = 1$ it would be the Manhattan. If $p = \infty$, then it would be maximum distance or Chebyshev [36].

2.5.TOOLS

Safe path planning application was developed with machine learning. In this thesis, many technologies such as Python, Javascript, Nodejs, Google Map API have been used while developing the application. The system infrastructure is shown in figure 2.10.

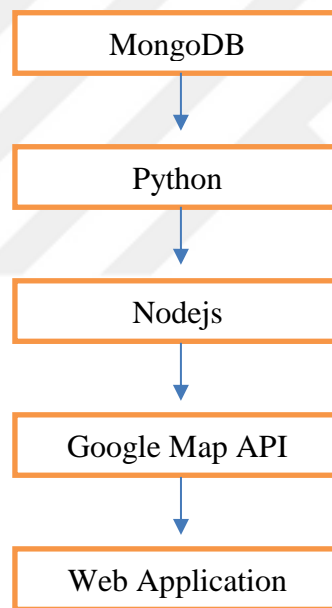


Figure 2.8: System architecture

2.5.1.Python

A widespread and strong interpreted language: Python. In contrast with R, It is a complete language and platform which can be utilized for both developing production systems and r&d. Furthermore, there are a lot of ML modules and libraries to select from, supplying multiple methods to accomplish each task. Optimal method to get started using Python for ML is to finalise a research.

- You will be forced to setup and begin the Python interpreter.
- You will be given a bird's eye view of how to step along a small research.
- You will be given confidence, even may continue to your own small researches.
- Python's core benefit is considered as a broad community backing and comprehensive set of ML libraries.
- One of the main benefits of Python is regarded as flexibility. Because of the option to prefer between Object-oriented programming's approach and scripting, it is proper for all aim.

In this application, the Python libraries numpy and pyclustering are used. The Pyclustering library is preferred to use clustering algorithms. Pyclustering is a Python data mining library which is focused on bio-inspired algorithms (based on oscillatory neural networks), and on classical algorithms that are widely used in many applications. The library strongly covers cluster analysis and oscillatory networks [37].

2.5.2. Google Map API

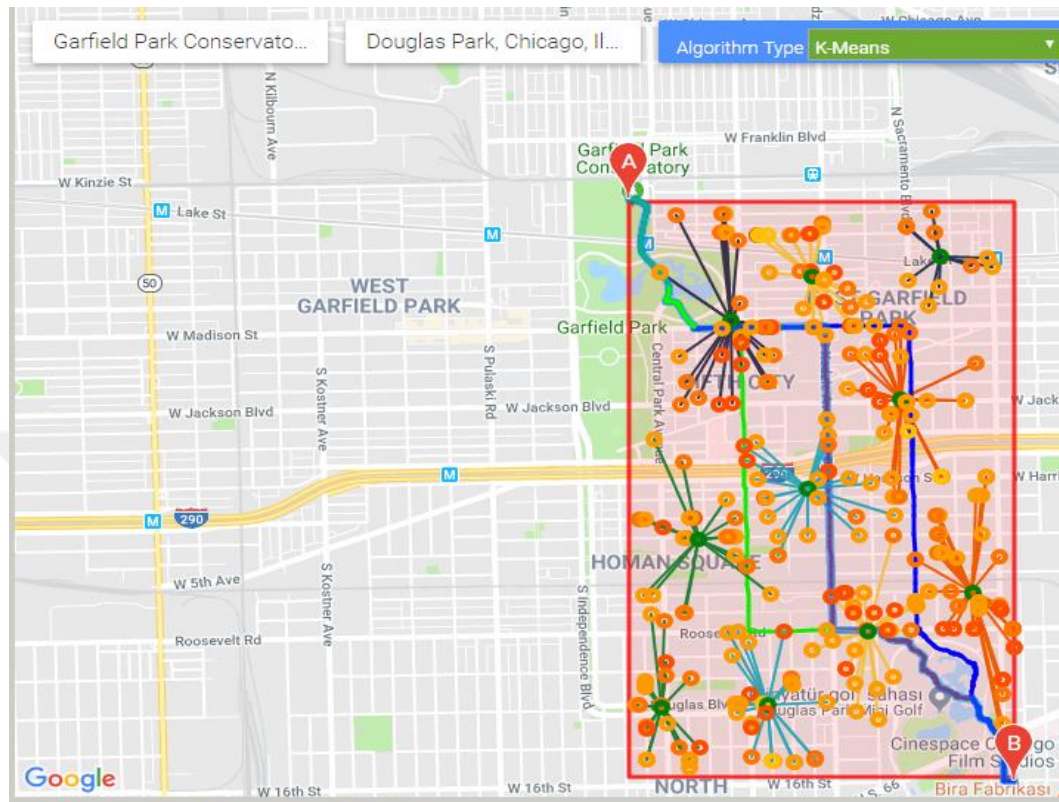
In the safe path planning, the Javascript library of the Google Map API is used to help draw maps and routes. The benefits of using the Google Map API are listed below.

- The Google Maps JavaScript API features provides which you can modify using layers and styles, controls and events, and various services and libraries [38].
- When planning a safe route, route by specifying the starting point (latitude / longitude) and the ending point (latitude / longitude).
- Polyline is drawn according to the route determined in the route planning and the scope of the crime scene is shown.
- Marking of crimes belonging to polyline plotted area with marker.

2.6.APPLICATION

An application has been developed, which is a web interface for safe path planning with machine learning. The crime data of the Chicago area have been studied. As indicated in Figure 2.11, one of the K-Means, K-Median, K-Medoid, X-Means, Expectation-Maximization algorithms is selected in the interface of the application while the road is planned. Then, the

starting point and destination points are selected. Safe route planning is performed on the map according to the determined road route and selected algorithm.



Red colored route is predicted as safest one

Figure 2.9: Safe path planning

For example: The starting point on the map is selected from Garfield Park, Chicago, and the destination point is selected from Douglas Park, Chicago.

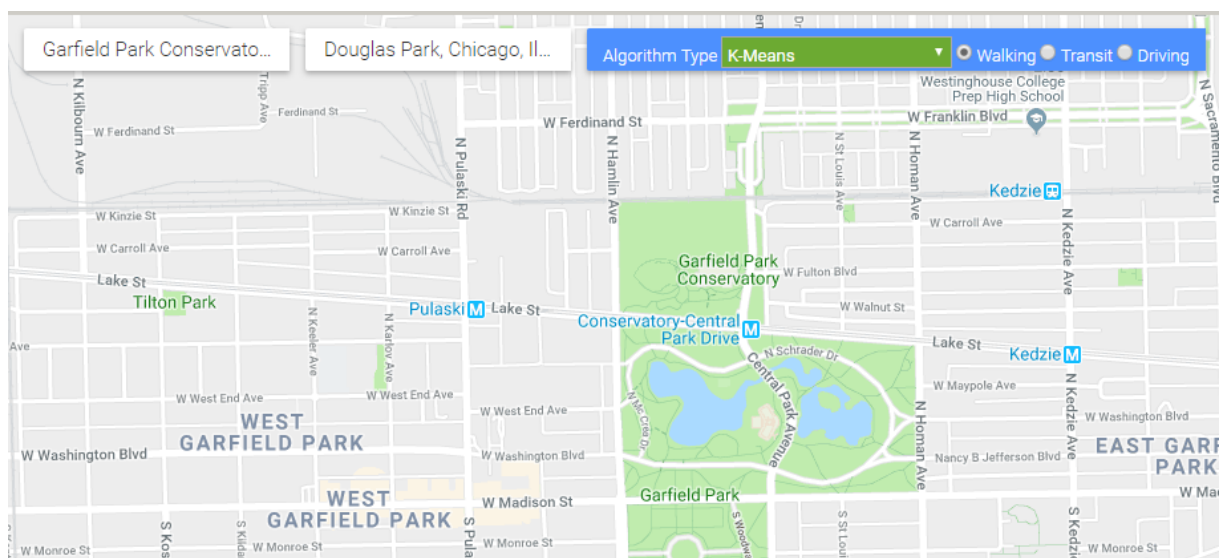


Figure 2.10: Example of safe path planning

Crimes in the area between the starting and ending points are shown on the map. The map also contains markers of crime places. The crimes are displayed in different colors on the map according to their value. We found three alternative routes according to the desired location (Red-Green-Blue Alternative Route).

Figure 2.12 shows the safe path planning by selecting the K-Means. Using the K-Means, crime data are grouped together. The distances of the center point of each cluster according to three different alternative roads are calculated. In addition, according to the crime values in each crime group, the arithmetic average of the crime is included in the account. As a result of each set of calculations, safe values are found for each path. The results show which of the three different routes is the safest route. The calculations in this example help to determine that the red colored path is the safest.

3. RESULTS

K-Means, K-Medians, K-Medoids, X-Means and Expectation Maximization were used in the clustering algorithms. The K-Medians method was calculated with K-Medoid algorithms and Euclidean, Manhattan, and Minkowsky methods as the distance measure. In this study, test results were obtained on a computer with Intel Core i7-3632QM CPU @ 2.20 GHZ processor, 6.00 GB RAM capacity and 64-bit operating system. In each clustering algorithm, every k value was calculated, and the result was repeated five times. The best safe values were selected from the five results obtained for each k value. Higher safe values correspond to higher confidence levels, according to which the safer road is determined.

The crime dataset of Chicago was used to obtain the test results in safe road planning. For the test results in practice, the starting point for the Chicago map was selected from Addams / Medill Park, Chicago, the arrival point was Garfield Park, and all test results were obtained accordingly.

We found the center point values (mean values) using the selected clustering algorithm. The mean lines are straight lines to show the crime that we found while we draw. As a result, the crime enables us to better understand the center point values (mean values). In addition, the average center point value (mean value) of crime is more secure. In the following visual map, the crimes in the selected place are shown using the clustering method. Figure 3.1 shows application of safe path planning.

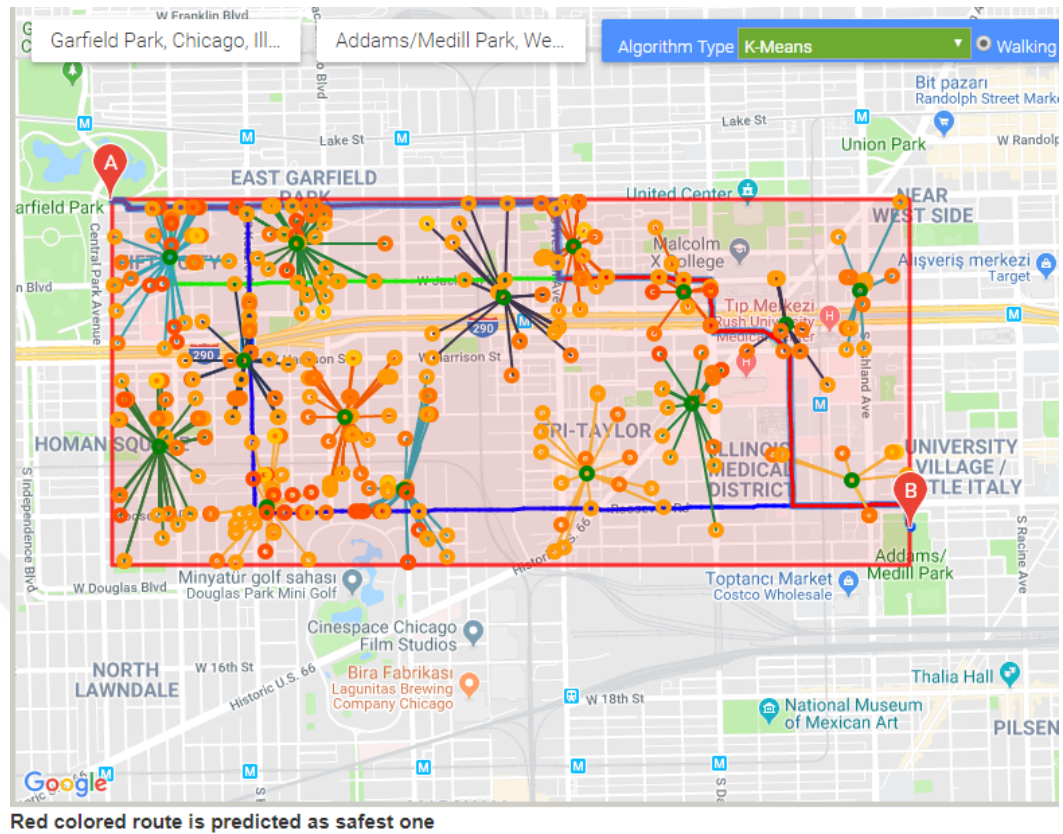


Figure 3.1: Application of safe path planning

3.1.CHOICE OF K-VALUE

The true selection of "k" is frequently uncertain, with interpretations based on the shape and scale of the allocation of points in a data set and also the demanded clustering resolution of the user. Moreover, if each data point is regarded its own cluster, rising k without punishment will usually decrease the sum of error in the resulting clustering, to the extreme case of no error. Then, heuristically, the optimal selection of k will set a balance between maximum compression of the data utilizing a single cluster, and maximum exactness by appointing each data point to its own cluster. When a suitable value of k is not explicit from previous knowledge of the properties of the data set, it must be selected in someway [39]. Furthermore, observing or visualizing the allocation of data points along groups is a beneficial and simple method to obtain insight into how the algorithm is separating the data for each k.

The k values are specified for use in all algorithms. The k values we use are 3, 4, 6, 8, 10, 12, 15, 20.

The meanings of the values used in the tables are as follows:

1. K-value: The letter "k" indicates the number of sets.
2. Safe values: “minimum distance” divided by mean value which is total of crime values assigned to that mean.
3. Min-Distances: Where “minimum distance” average of min distances from route to means.
4. Route: Represents the safe route determined as the result of the algorithm. (Red, Green, Blue)

The K-value, safe values, min-distances, and route values for the K-Means clustering algorithm are indicated in Table 3.1.

Table 3.1: K-Means

K-value	Safe values	Min-Distances	Route
3	0.636	0.105	green
4	0.837	0.112	green
6	1.24	0.151	blue
8	1.21	0.090	green
10	1.520	0.128	blue
12	1.79	0.081	green
15	2.20	0.096	green
20	2.28	0.101	green

The K-value, safe values, min-distances, and route for the Euclidean distance measure used with the K-Median clustering algorithm are shown in Table 3.2.

Table 3.2: K-Median & Euclidean distance

K-value	Safe values	Min-Distances	Route
3	0.242	0.174	red
4	0.368	0.134	blue
6	0,790	0.149	green
8	0.664	0.142	blue
10	0.808	0.157	blue
12	1.27	0.129	blue
15	1.96	0.128	blue
20	1.98	0.149	blue

The K-value, safe values, min-distances, and route for the Manhattan distance measure used with the K-Median clustering algorithm are shown in Table 3.3.

Table 3.3: K-Median & Manhattan distance

K-value	Safe values	Min-Distances	Route
3	0.232	0.159	red
4	0.361	0.148	blue
6	0.550	0.143	green
8	0.610	0.104	red
10	0.706	0.131	blue
12	0.860	0.111	green
15	1.46	0.087	red
20	1.79	0.0985	red

The K-value, safe values, min-distances, and route for the Miskowsky distance measure used with the K-Median clustering algorithm are shown in Table 3.4.

Table 3.4: K-Median & Minkowsky distance

K-value	Safe values	Min-Distances	Route
3	0.253	0.147	green
4	0.427	0.174	red
6	0.533	0.143	green
8	1.30	0.171	blue
10	0.947	0.137	blue
12	0.951	0.149	blue
15	1.01	0.104	blue
20	1.55	0.111	red

The K-value, safe values, min-distances, and route for the Euclidean distance measure used with the K-Medoid clustering algorithm are shown in Table 3.5.

Table 3.5: K-Medoid & Euclidean distance

K-value	Safe values	Min-Distances	Route
3	0.455	0.178	blue
4	0.507	0.116	blue
6	0.328	0.092	blue
8	0.629	0.056	blue
10	0.871	0.050	red
12	0.756	0.100	blue
15	0.571	0.071	blue
20	0.615	0.040	blue

The K-value, safe values, min-distances, and route for the Manhattan distance measure used with the K-Medoid clustering algorithm are shown in Table 3.6.

Table 3.6: K- Medoid & Manhattan distance

K-value	Safe values	Min-Distances	Route
3	0.330	0.087	green
4	0.958	0.119	green
6	0.536	0.150	blue
8	0.596	0.130	red
10	0.970	0.131	blue
12	1.27	0.109	green
15	1.10	0.090	red
20	1.32	0.109	blue

The K-value, safe values, min-distances, and route for the Miskowsky distance measure used with the K-Medoid clustering algorithm are shown in Table 3.7.

Table 3.7: K- Medoid & Minkowsky distance

K-value	Safe values	Min-Distances	Route
3	0.242	0.084	red
4	0.658	0.079	blue
6	0.479	0.015	green
8	1.65	0.151	blue
10	1.32	0.083	red
12	1.97	0.146	green
15	2.50	0.083	green
20	2.17	0.104	red

The K-value, safe values, min-distances, and route values for the X-Means clustering algorithm are shown in Table 3.8.

Table 3.8: X-Means

K-value	Safe values	Min-Distances	Route
3	0.310	0.112	red
4	0.310	0.112	red
6	0.829	0.095	green
8	2.23	0.129	green
10	2.23	0.129	green
12	1.37	0.148	green
15	1.17	0.100	red
20	1.29	0.092	green

The K-value, safe values, min-distances, and route values for the Expectation-Maximization clustering algorithm are shown in Table 3.9.

Table 3.9: Expectation-Maximization

K-value	Safe values	Min-Distances	Route
3	0.130	0.159	green
4	2.10	0.050	blue
6	2.09	0.082	red
8	2.10	0.050	blue
10	1.79	0.056	red
12	1.10	0.056	green
15	5.97	0.1201	red
20	5.73	0.0769	green

3.2.SIMULATION

We have simulated travelling by foot through all the tested of routes evaluating all the distances from every step to crime point and amount of crimes for each step. So the average of distances and crime amount to step points were calculated.

Simulation gave us following results:

1. Start of Simulation, Route: Red

Number of crimes close to step: 1256

Total steps: 202

Avarage number of crimes to every step: 6.21

Avarage distance to crimes: 0.11

2. Start of Simulation, Route: Greem

Number of crimes close to step: 1122

Total steps: 202

Avarage number of crimes to every step: 5.55

Avarage distance to crimes: 0.12

3. Start of Simulation, Route: Blue

Number of crimes close to step: 1387

Total steps: 196

Avarage number of crimes to every step: 7.07

Avarage distance to crimes: 0.12

4. DISCUSSION

In recent years, the increasing crime rate has become a serious problem in many countries. Companies and governments worldwide attempt to use artificial intelligence for crime reduction, prevention, and response to ongoing crime. Most studies are based on the idea that crimes are foreseeable. The analysis of major crime figures was technologically impossible a few decades ago but is now enabled by the latest developments in machine learning. There have been many studies on these issues. In general, crime analysis, crime prediction, patrol planning are made for law enforcement officers, and crimes are predicted by analyzing social media data or news sites.

In this study, a different solution is presented from other studies. We aim to ensure that passengers travel safer when they travel to a region. In this safe journey, alternative routes are provided according to the road safety ratios between the starting and arrival points, which are determined by the passenger on the map, so it is possible to determine the more secure route. Thus, we help passengers to travel more safely from one place to another.

In this research, the clustering algorithms used are K-Means, K-Median, K-Medoid, X-Means, Expectation-Maximization.

The first method is the K-Means algorithm. Table 3.1 shows the results obtained from the K-Means algorithm. Different results were obtained according to K-value (number of clusters). The safe value was lower when the number of clusters was small. But when k-value increases, the confidence rate also increases. We can draw a conclusion from here. As the K value increases, the safe value, that is, the confidence rate, increases. The k-value for the safest way in the test results was 20 while the safe value was 2.28. According to the result, it is determined that the green route, which is one of the three different routes, is safer.

As a second method, Euclidean distance is used together with K-Median algorithm. Table 3.2 shows the results obtained from K-Median & Euclidean distance. As the K value increases, the safe value, that is, the confidence rate, increases. The k-value for the safest way in the test results was 20 while the safe value was 1.98. According to the result obtained, it is determined that the blue route, which is one of the three different routes, is safer.

In the same algorithm, Manhattan, which is a different distance measure method, is used. Table 3.3 shows the results obtained from K-Median & Manhattan distance. Again as Euclidean distance, as the K value increases, the safe value, that is the confidence rate, is found to increase. The k-value for the safest way in the test results was 20 while the safe value was 1.79. According to the result obtained, it is determined that the red route which is one of the three different routes is safer.

The last Minkowsky distance was applied together with the K-Median algorithm. Table 3.4 shows the results obtained from K-Median & Minkowsky distance. As in the other distance measure methods, it has been found that the safe value increases as the K value increases. The k-value for the safest way in the test results was 20 while the safe value was 1.55. According to the result obtained, it is determined that the red route which is one of the three different routes is safer.

As a third method, the Euclidean distance used in conjunction with the K-Medoid algorithm is used. Table 3.5 shows the results obtained from K-Medoid & Euclidean distance. As the K value increases, the safe value, that is, the confidence rate, increases. The k-value for the safest way is 20, while the safe value is 0.615. According to the result, it is found that the blue route which is one of the three different routes is safer.

In the same algorithm, Manhattan, which is a different distance measure method, is used. Table 3.6 shows the results obtained from K-Medoid & Manhattan distance. Again as Euclidean distance, as the K value increases, the safe value, that is the confidence rate, is found to increase. The k-value for the safest way in the test results was 20 while the safe value was 1.32. According to the result obtained, it is determined that the blue route, which is one of the three different routes, is safer.

The last Minkowsky distance was applied with the K-Medoid algorithm. Table 3.7 shows the results obtained from K-Medoid & Minkowsky distance. As in the other distance measure methods, it has been found that the safe value increases as the K value increases. The k-value for the safest way in the test results was 20 while the safe value was 2.17. According to the result obtained, it is determined that the red route which is one of the three different routes is safer.

As the fourth method, X-Means algorithm is used. Table 3.8 shows the results obtained from X-Means. For the safest way, k values were 8 and 20, while safe value was 2.23. According to the result obtained, it is determined that the green route, which is one of the three different routes, is safer.

As the fifth method, Expectation-Maximization clustering algorithm is used. Table 3.9 shows the results obtained from Expectation-Maximization. Each k-value has different results in the EM algorithm. Many k-values have produced very successful results.



5. CONCLUSION AND RECOMMENDATIONS

5.1.CONCLUSION

This final chapter concludes the thesis and makes suggestions for future implementation study.

Inferences and comparisons were made from the results of clustering algorithms for safe path planning. In the K-Means algorithm, when the K value increases, the safe value, which is the confidence rate, increases. In addition, the safe values of many k values produce successful results. When the K-Median algorithm is used with the Euclidean distance, it yields higher safe values than other distance measures. The K-Median with Manhattan and Miskowsky distance measurements has similar safe values.

For K-Medoids, when the Euclidean distance is used, lower safe values are obtained compared to other distance measures. When the Minkowsky distance is used with K-Medoids, more successful safe values are obtained than other distance measures. In X-Means algorithms, unlike other algorithms, the safe value does not increase when the k value increases. Good results are obtained at k values of 8 and 10 in the tests.

In the Expectation Maximization algorithm, different k values are obtained. Unlike the methods in other algorithms, a random k value is determined for a maximum number of sets, which makes it difficult to show the desired k value. However, many k values have produced notably successful results.

In this study, successful results have been obtained using the Expectation Maximization, K-Medoids with Minkowsky distance, and K-Means algorithms, as shown in the comparison.

5.2.RECOMMENDATIONS

This research now working only in walking mode and in walking mode shows us safe path planning with alternatives. As an improvement in future we can add the information about traffic, car crash, street construction work, weather condition information, meetings etc. to this application, after adding this type of information should be added driving mode. For finding safe path planning we used K-Means, K-Median, K-Medoid, X-Means, Expectation-

Maximization algorithm and another algorithms should be added according to functions. We will improve speed and fixed error ratio by passing various performance tests.



REFERENCES

- [1]. *Chicago Data Portal*, <https://data.cityofchicago.org/Public-Safety/Crimes-2018/3i3m-jwuy>, [Visit Date: February 5, 2018].
- [2]. *Timeline of ML*, https://en.wikipedia.org/wiki/Timeline_of_machine_learning, [Date of visit: March 17, 2018].
- [3]. Gonzalez, V., 2018, *A Brief History of Machine Learning*, <http://www.synergicpartners.com/en/espanol-una-breve-historia-del-machine-learning>, [Date of visit: April 3, 2018].
- [4]. Gezer, M., ve Saylan, S., Aralık 2017, *Açık Kaynaklı Makine Öğrenmesi Kütüphaneleri, Mühendislikte Yapay Zekâ ve Uygulamaları*, Torkul, O. ve Gülseçen S. (eds.), 4. Bölüm, Sakarya Üniversitesi Kütüphanesi Yayınevi, Sakarya, ISBN: 978-605-4735-98-3, 55-56.
- [5]. Copeland, M., 2016, *What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?*, <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>
- [6]. NG, A., *Machine Learning*, <https://www.coursera.org/learn/machine-learning>, [Date of visit: April 14, 2018].
- [7]. Pyle, D. and San Jose, C., 2015, *An executive's guide to machine learning*, <https://www.mckinsey.com/industries/high-tech/our-insights/an-executives-guide-to-machine-learning>, [Date of visit: April 16, 2018].
- [8]. *Types of Machine Learning Algorithms You Should Know*, <https://towardsdatascience.com/types-of-machine-learning-algorithms-you-should-know-953a08248861>, [Date of visit: April 16, 2018].
- [9]. *Machine Learning: What it is and Why it Matters*, <https://www.simplilearn.com/what-is-machine-learning-and-why-it-matters-article>, [Date of visit: April 17, 2018].
- [10]. *Machine Learning, Part II: Supervised and Unsupervised Learning*, http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm, [Date of visit: April 21, 2018].
- [11]. *A Tutorial on Clustering Algorithms*, https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/, [Date of visit: April 21, 2018].
- [12]. Maini, V. and Sabri, S., 2017, *Dimensionality reduction*, *Machine Learning for Humans*, 61.
- [13]. *Introduction to Dimensionality Reduction*, <https://www.geeksforgeeks.org/dimensionality-reduction/>, [Date of visit: May 1, 2018].
- [14]. Ayodele, T., 2010, *Types of Machine Learning Algorithms*, *New Advances in Machine Learning*, Zhang, Y. (eds.), InTech, China, ISBN 978-953-307-034-6, 20-21.

- [15]. Barnadas, M.V., 2016, *Machine Learning Applied To Crime Prediction*, A Degree Thesis Submitted to the Faculty of the Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona Universitat Politècnica de Catalunya, Barcelona.
- [16]. Chen, X., Cho, Y., and other, 2015, Crime Prediction Using Twitter Sentiment and Weather, *2015 IEEE Systems and Information Engineering Design Symposium*.
- [17]. Iqbal, R., Murad, M., Mustapha, A., 2013, An Experimental Study of Classification Algorithms for Crime Prediction, *Indian Journal of Science and Technology - March 2013 Indian Journal of Science and Technology*, Universiti Putra Malaysia, ISSN: 0974-6846.
- [18]. Shama, N., 2017, *A Machine Learning Approach to Predict Crime Using Time and Location Data*, Department of Computer Science & Engineering / BRAC University.
- [19]. Kianmehr, K., Reda Alhajj, R., 2006, *Crime Hot-Spots Prediction Using Support Vector Machine*, 2006 IEEE.
- [20]. Yu, C. H., Ward, M. W., and other, 2011, Crime Forecasting Using Data Mining Techniques, *11th IEEE International Conference on Data Mining Workshops*, 779-786.
- [21]. MacQueen, J., 1967, Some methods for classification and analysis of multi-variate observations, *In: Proc. of the Fifth Berkeley Symp. on Math., Statistics and Probability*, LeCam, L.M., and Neyman, J., (eds.), Berkeley: U. California Press, 281.
- [22]. Berry, M., Linoff, D., 2004, *Data Mining Techniques*, Wiley Publishing Inc.
- [23]. *K-means Clustering*, <http://ba-finance-2013.blogspot.com/2012/09/k-means-clustering.html>, [Date of visit: May 5, 2018].
- [24]. Kanungo, T., Mount, D., M, Netanyahu, N., S., Piatko, C., Silverman, R., Wu, A. Y., 2006, *The Analysis of a Simple k-means Clustering Algorithm*, ACM.
- [25]. Anderson, B., Gross, D., Musicant, D., and other, 2006, Adapting K-Medians to Generate Normalized Cluster Centers, *Proceedings of the Sixth SIAM International Conference on Data Mining*, 165-175.
- [26]. *K-medoid Clustering*, https://www.researchgate.net/figure/k-medoids-clustering_fig4_282897075, [Date of visit: May 6, 2018].
- [27]. Pelleg, D., Moore A.W., 2018, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *Proceedings of the Seventeenth International Conference on Machine Learning*.
- [28]. Schwarz, Gideon E., 1978, *Estimating the dimension of a model*, *Annals of Statistics*, 461-464.
- [29]. Ishioka T., 2005, An Expansion of X-means for Automatically Determining the Optimal

Number of Clusters, *Proc. of The 4th IASTED International Conference on Computational Intelligence*, 91-96.

- [30]. Dempster, A.P., Laird, N.M., Rubin, D.B., 1977, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*. 39 (1): 1–38.
- [31]. *Expectation maximization*, https://www.projectrhea.org/rhea/index.php/Expectation_Maximization_Old_Kiwi, [Date of visit: May 6, 2018].
- [32]. Dinler, M., 2014, *Kümeleme Analizi Yöntemlerinin Hayvancılık Verilerinde Karşılaştırılmalı Olarak İncelenmesi*, Yüksek Lisans Tezi, Bingöl Üniversitesi Fen Bilimleri Enstitüsü.
- [33]. Demiralay, M., Çamurcu, Y., 2005, Agnes ve K-means Algoritmalarındaki Kümeleme Yeteneklerinin Karşılaştırılması, *İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi*, 8(2): 1-18.
- [34]. *What are some best use cases to use Chebyshev or Manhattan distances if there are any?*, <https://www.quora.com/What-are-some-best-use-cases-to-use-Chebyshev-or-Manhattan-distances-if-there-are-any>, [Date of visit: May 8, 2018].
- [35]. Atbaş, A.C.G., 2008, *Kümeleme Analizinde Küme Sayısının Belirlenmesi Üzerine Bir Çalışma*, Yüksek Lisans Tezi, Ankara Üniversitesi, Fen Bilimleri Enstitüsü.
- [36]. Cordeiro de Amorim, R. and Mirkin, B., 2012, *Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering*, *Pattern Recognition*, no. 3, 1061–1075.
- [37]. Novikov, A., 2017, *Pyclustering library*, <https://github.com/annoviko/pyclustering>, [Visit Date: April 25, 2018].
- [38]. Google Map API, <https://developers.google.com/maps/documentation/javascript/tutorial>, [Visit Date: May 3, 2018].
- [39]. *Determining the number of clusters in a data set*, https://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set, [Visit Date: April 19, 2018].

CURRICULUM VITAE

Personal Information	
Name Surname	Yasin Uğur BELEK
Place of Birth	İstanbul
Date of Birth	09.05.1989
Nationality	<input checked="" type="checkbox"/> T.C. <input type="checkbox"/> Other:
Phone Number	0551 187 54 51
Email	yasinugurb@gmail.com
Web Page	https://www.linkedin.com/in/yasinugurbelek



Educational Information	
B. Sc.	
University	Anadolu University
Faculty	Faculty of Economics
Department	Public administration
Graduation Year	01.01.2013

M. Sc.	
University	İstanbul University
Institute	Institute of Graduate Studies in Science and Engineering
Department	Department of Computer Engineering
Programme	Computer Engineering Programme