Avrupa Bilim ve Teknoloji Dergisi
Sayı 38, S. 296-310, Ağustos 2022

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 38, pp. 296-310, August 2022

Research Article

Sampling Techniques and Application in Machine Learning in order to Analyse Crime Dataset

Ayla Sayli[1*], Sevil Basarir[2]

[1] Yildiz Technical University, Faculty of Chemistry and Metallurgical,Department of Mathematical Engineering, Istanbul, Turkey, (ORCID: 0000-0003-0409-537X), sayli@yildiz.edu.tr

[2*] Yildiz Technical University, Faculty of Chemistry and Metallurgical,Department of Mathematical Engineering, Istanbul, Turkey, (ORCID: 0000-0002-1599-0727), basarirsevil@gmail.com

(First received 11 May 2022 and in final form 14 June 2022)

(**DOI:** 10.31590/ejosat.1115323)

ATIF/REFERENCE: Sayli, A. & Basarir, S. (2022). Sampling Techniques and Application in Machine Learning in order to Analyse Crime Dataset. *European Journal of Science and Technology,* (38), 296-310.

Abstract

Machine learning enables machines to learn information and make inferences using the information it has learned. In this article, five years of crime data were analyzed and the learning process was completed with the data in the machine's hands. One-Hot Encoding and Min-Max Normalization methods and Principal Component Analysis algorithm were used in the analysis of the data. The model was asked to predict whether the criminal could be caught, the security of the area, and the type of crime committed using the K-Nearest Neighborhood, Random Forest and Extreme Gradient Boosting algorithms. However, no matter how successful the model is in imbalanced datasets, the result will be misleading. Therefore, the main purpose of this article is to transform the imbalanced data into a balanced one by various methods and to find the most accurate sampling method for the data, which is compatible with the classification method. For this purpose, one statistical sampling method (Stratify), three over sampling method (Random Over Sampler, Synthetic Minority Over, Adaptive Synthetic), three under sampling method (Random Under Sampler, Near Miss, Neighborhood Cleaning Rule) and mix samplig method (Smote Tomek) have been applied to avoid imbalance of data in target areas such as Arrest, Crime Type,Security. As a result of the sampling methods applied, efficient and effective results were obtained.

Keywords: Sampling Techniques, Classification, Data Pre-Processing, Machine Learning, Crime Analysis, Data Analysis, Data Visualization.

Suç Veri Setini Analiz Etmek İçin Makine Öğreniminde Örnekleme Teknikleri ve Uygulaması

Öz

Makine öğrenmesi, makinelerin bilgiyi öğrenmesini ve öğrendiği bilgiyi kullarak çıkarımlar yapmasını sağlar. Bu makalede, beş yıla ait suç verileri ele alınarak analiz edildi ve makinenin elindeki verilerle öğreme işleminin tamamlanması sağlandı. Verinin analizi sürecinde One-Hot Encoding ve Min-Max Normalizasyon methodları ile Principal Component Analysis algoritması kullanıldı. Modelden suçlunun yakalanıp yakalanamaması, bölgenin güvenliği ve işlenen suçun tipini K-Nearest Neighborhood, Random Forest ve Extreme Gradient Boosting algoritmaları kullanılarak tahmin etmesi istendi. Fakat dengesiz veri setlerinde model ne kadar başarılı olursa olsun sonuç yanıltıcı olur. Bu nedenle bu makalenin asıl amacı dengesiz verinin çeşitli methodlarla dengeli hale dönüştürülmesi ve veri için sınıflandırma methodu ile uyumlu en doğru örnekleme methodunu bulmaktır. Bu amaçla tutuklanma, suç tipi, güvenlik gibi hedef alanlarında verinin dengesizliğinin önüne geçmek için bir tane istatistiki örnekleme methodu (Tabakalaştırma), üç tane üst önekleyici method (Rastgele Üst Örnekleyici, Sentetik Azınlık Üstü, Uyarlamalı Sentetik), üç tanem alt örnekleyici method (Rastgele Alt Örnekleyici, Ramak Kala, Yakın Komşu Temizleme Kuralı) ve bir tane alt ve üst karışık örnekleme methodu (Smote Tomek) uygulanmıştır. Uygulanan örnekleme yöntemleri sonucunda verimli ve etkili sonuçlar elde edilmiştir.

Anahtar Kelimeler: Örnekleme Teknikleri, Sınıflandırma, Veri Ön İnceleme, Makine Öğrenmesi, Suç Analizi, Veri Analizi, Veri Görselleştirme.

* Corresponding Author: sayli@yildiz.edu.tr

1. Introduction

Machine learning is an extremely popular topic among today's technologies. A lot of contributions are made to the literature with assorted studies conducted day by day. Machine learning is basically divided into three main titles as supervised, unsupervised and reinforced learning. Supervised learning consists of Regression and Classification studies. In the fields of machine learning and statistics, the classification problem is to find which of a set of categories a new observation belongs to, using a run set of basic observations and known categories.

Sampling techniques is one of sub-fractions of /classification studies . The sampling method is used to make the imbalanced classes balanced and to choose the right sample for the algorithms. The success of the classification study will be very misleading when the correct sample is not selected or a sample with a majority of samples belonging to a certain class is selected. In order to prevent this, the process should be continued by making the imbalanced data set balanced by using various sampling techniques, before starting to work on classification algorithms on imbalanced data sets.

Sampling techniques can be applied according to statistical methods, over sampling methods, under sampling methods or mix sampling methods. There are a lot of articles in the literature about sampling methods in machine learning. Most recent articles related this topic are given in Table 1.

Table 1. Literature Studies of Sampling Techniques

Publication	Method	Summary
A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data (Batista, G. E., Prati, R. C., & Monard, M. C., 2004)	Random over-sampling, Random under-sampling, Tomek links, Condensed Nearest Neighbor Rule, One-sided selection, CNN + Tomek links, Neighborhood Cleaning Rule, Smote, Smote + Tomek links, Smote + ENN	In general, oversampling methods gave better results. Smote + Tomek and Smote + ENN achieved higher success than other algorithms.
Common Survey Sampling Techniques (Hibberts, M., Burke Johnson, R., & Hudson, K., 2012)	Simple random sampling, Systematic sampling, Clustered sampling, Convenience sampling, Quota sampling, Purposive sampling, Referral sampling, Network sampling, Snowball sampling, Nonrandom sampling, Referral sampling	The relation of sampling algorithms used with each other was compared.
Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset (Junsomboon, N., & Phienthrakul, T., 2017)	SMOTE, Neighborhood Cleaning Rule, Naïve Bayes, SMO, KNN	NCL+SMOTE method has been compared with all methods used and it has been determined that it gives the best results.
Comparison of The Different Sampling Techniques For Imbalanced Classification Problems in Machine Learning (Zhihao, P., Fenglong, Y., & Xucheng, L. ,2019)	Random Over Sampling, Random Under Sampling, SMOTE, Kmeans Clustering, XG Boost	When the SMOTE and MSMOTE technique was used with Gradient Boostind and XG Boost, it gave better results than the others.
Effect of machine learning re-sampling techniques for imbalanced datasets in F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients (Xie, C., Du, R., Ho, J. W., Pang, H. H., Chiu, K. W., Lee, E. Y., & Vardhanabhuti, V., 2020)	ADASYN, SMOTE, borderline-SMOTE, random undersampling (RUS), NearMiss, Tomek link (TL), edited dataset using nearest neighbours (ENN), SMOTE-TL, SMOTE-ENN, LR, SVM, RF, XGBoost	The best result was achieved by RF classifier and SMOTE method.

Although crime is not systematic or random, it has existed since the ages of humanity. In the face of increasing crime rates in recent years, even the smallest clues about crime prevention or prevention are especially important for law enforcement. Crime analysis; it can be used to examine issues such as location, time, type of crime, the way the crime was committed, air temperature, season, population of the region where the crime was committed, crime trends and to determine their relationship with each other. For this reason, studies on crime analysis and the outputs of these studies are particularly important. Studies on the crime data set are given in table 2.

Table 2. Literature Studies on Crime Dataset

Publication	Method	Summary
Crime Analysis and Prediction Using Data Mining (Sathyadevan, S., Devan, M. S., & Gangadharan, S. S., 2014)	K-Nearest Neighbour (KNN), Boosted Decision Tree	Although it gives better results than the boosted decision tree knn algorithm, the success rate is generally low.
Survey of Analysis of Crime Detection Techniques Using Data Mining and Machine Learning (Prabakaran, S., & Mitra, S., 2018)	Genetic algorithm, Hidden Markov Model(HMM), Naive Bayesian, Cumulative logistics model, K-mean Clustering, K-mode Clustering, Neural Network, Kernel density estimation, Logistic regression, Random forest algorithm, Influenced association rule, J48 algorithm.	Different crime detection techniques have been studied according to each crime type. Mathematical and verbal techniques are mentioned and related studies in the literature are given.
Predicting the Type of Crime: Intelligence Gathering and Crime Analysis (Albahli, S., Alsaqabi, A., Aldhubayi, F., Rauf, H. T., Arif, M., & Mohammed, M. A., 2021)	PCA, FAMD, Naive Bays, Random Forest, KNN, Decision Tree	The naive bayes algorithm was determined as the most appropriate data set for the relevant data set.

There are assorted studies on sampling methods and crime predictions in the literature. However, the effects of sampling techniques were not examined while analyzing and estimating the crime data set. Here, unlike previous studies, we will examine the effect of sampling techniques on the success from four distinct perspective (statistical, over, under, mix sampling) on the crime data set.

Firstly stabilize the imbalanced data set by using various sampling methods, and then we will measure the success of the model by subjecting the balanced data set to classification methods. By comparing different sampling methods with various classification methods, we will suggest the most suitable sampling method for our model.

The article consists of a total of six chapters. In the first part, there is a general-purpose subject and literature summary. In the second part, the materials and methods are mentioned. Its subtitles consist of sampling methods, classification algorithms and evaluation metrics. In the third section, how the data was collected and the data set in the third part includes reviews on crime data. First, how the data is collected, what is target, how it is processed, which methods are applied on the data, and finally, the visualization of the data is mentioned. In the fourth chapter, the experimental setup stages are explained. The sixth chapter consists of the results of all studies and comparative evaluations.

2. Material and Method

2.1. Sampling Techniques

Sampling is a popular approach used to eliminate class imbalance. The purpose of sampling methods is to generate data with a relatively more balanced class distribution. Thus, classification algorithms can better capture the decision boundary between majority and minority classes. (Kurin, S., Steinshamn, S. I., & Saerens, M., 2017)

Sampling can be done by using methods based on statistical techniques, or by over sampling to complete the majority with multiplication operations in the equating of classes, or by using under sampling techniques to complete the majority with education operations in the equating of classes. In this article, Stratified from statistical methods,

Random Over Sampler, SMOTE, ADASYN from over sampling methods (Random Over Sampler, SMOTE, SMOTENC, SMOTEN, ADASYN, Borderline SMOTE, KMeans SMOTE, SVM SMOTE), Neighbourhood Cleaning Rule, Random Under Sampler, Near Miss from under sampling methods (Cluster Centroids, Condensed Nearest Neighbour, Edited Nearest Neighbours, Repeated Edited Nearest Neighbours, AllKNN, Instance Hardness Threshold, Near Miss, Neighbourhood Cleaning Rule, One Sided Selection, Random Under Sampler, Tomek Links) and Smote Tomek from combine sampling methods (Smote Tomek, Smote Enn) are used.

2.1.1. Stratified Sampling (SS)

In statistical science, stratified sampling is the acquisition of data from a population with a special shape probability sampling method. The feature that distinguishes the stratified sampling method from other probability sampling methods is that all the elements in the population consist of several groups and strata that are similar to each other according to certain characteristics. The layer elements are similar to each other but very distinctly different from the other layer elements. In the layer example, the sample elements are selected such that there are representatives in the sample for each population layer. (Meng, X., 2013)

2.1.2. Random Over Sampling (ROS)

Random Oversampling is one of the oldest methods and has proven to be quite consistent in its results. (Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2002). Basically, it aims to equate classes by selecting and copying random elements from the minority class and repeating them until they match the number of elements of the majority class.

2.1.3. Random Under Sampling (RUS)

The Random Under Sampler method is one of the oldest techniques used to remove imbalances in the data set. It is basically based on the principle of equating the minority class with the majority class by removing random elements from the

majority class. But it can increase the variance of the classifier and there is a risk of discarding important samples. (Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V., 2018)

2.1.4. Synthetic Minority Over Sampling (SMOTE)

SMOTE (Synthetic Minority Over-Sampling Technique) is an oversampling process that produces synthetic data. The main idea of the method is to create new instances of the minority class by performing certain operations between instances of the minority class. Synthetic samples are produced as follows: the difference between the examined feature vector (E_i) and its nearest neighbor is taken, this difference is multiplied by a random number (δ) between 0 and 1, the final result is added to the examined feature vector and a new sample is formed. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., 2002).

$$E_{new} = E_i + (E_i - E_j)\delta$$

2.1.5. NearMiss Sampling (NMS)

The Near Miss sampling method aims to balance the class distribution by randomly eliminating majority class samples. First, the Euclidean distance between the majority and the minority class is calculated. Then, n samples of the majority class with the smallest distances to those in the minority class are selected. If the minority class has k instances, the closest method will result in k*n instances of the majority class. (DURAHİM, A. O.,2016). There are three variations of applying the Near Miss Algorithm to find the n closest instances in the Majority class:

- First method: Samples of the majority class are selected where the mean distances to the nearest k samples of the minority class are the smallest.
- Second method: Samples of the majority class with the smallest mean distances to the k-farthest k samples of the minority class are selected.
- Third method: It works in two steps. First, for each instance of minority class, M nearest neighbor is stored. Finally, majority class instances are selected for which the mean distance to N nearest neighbors is the largest.

2.1.6. Neighbourhood Cleaning Rule Sampling (NCR)

The Neighbor Cleaning Rule (NCL) is an under sampling method proposed in 2001. (Laurikkala, J.,2001) For two-class problems, he recommends a more comprehensive version of the Edited Nearest Neighborhood method. According to the working logic, first of all, the data has k close neighbors. (The default is three.) Then, if the selected data belongs to the majority class and there is a prediction error about its nearest neighbors, the selected data is removed. But if the selected data belongs to the minority class and there is an estimation error about its nearest neighbors, then the nearest neighbors belonging to the majority class are removed.

2.1.7. Adaptive Synthetic Sampling (ADASYN)

The ADASYN algorithm is an over sampling method and works by generating enough synthetic alternatives for each observation belonging to the minority class. How many data it needs to generate depends on how complicated it is to learn the original observation. It is just as difficult to learn an observation from the minority class, especially if there are multiple examples of observations with similar characteristics in the majority class. (He, H., Bai, Y., Garcia, E. A., & Li, S., 2008)

2.1.8. Smote Tomek Sampling (ST)

Tomek links are an under sampling method that identifies all pairs of data points that belong to different classes but are closest to each other. These data pairs defined by tomek links are called tomek links. Tomek Links is subject to two main conditions. First, the two selected data are the closest neighbors to each other. Second, these two data belong to different classes. Tomek connections are at the boundary of separation of the two classes. Therefore, removing the majority class from tomek connections increases class separation and thus reduces the number of instances of the majority class. (Tomek, I., 1976)

Smote, on the other hand, is an extreme learning technique and creates new minority class synthetic examples.

In the Smote-Tomek technique, on the other hand, to transform the imbalanced data into a balanced one, first, new minority class synthetic samples are created with the smote technique. Later, Tomek links are used to extract samples near the boundary of the two classes to increase the separation between the two classes. (Batista, G. E., Bazzan, A. L., & Monard, M. C.,2003)

2.2. Classification Algorithms

Machine learning algorithms are snippets of code that help people discover, analyze, and find meaning in complex datasets. Each algorithm creates the paths a machine will follow to achieve a specific goal. The goal of a machine learning model is to create or discover patterns that humans can use to make predictions or categorize information. Different algorithms analyze data differently. These are grouped by the machine learning techniques in which they are used: supervised learning, unsupervised learning, and reinforcement learning. The most widely used algorithms use regression and classification to predict target categories, find unusual data points, estimate values, and find similarities. Classification algorithms use predictive calculations to assign data to preset categories such as Support Vector Machine, K-Nearst Neighborhood, Logistic Regression, Desicion Tree, Random Forest, A Neural Network, Extreme Gradient Boostine, Naive Bayes. In this studies, for testing sampling techniques it has been used K-Nearst Neighborhood, , Random Forest and Extreme Gradient Boostine algorithms.

2.2.1. K-Nearest Neighborhood Algoritm (KNN)

The k-nearest neighbor (KNN) algorithm is one of the supervised learning algorithms that is easy to implement. Although it is used in solving both classification and regression problems, it is mostly used in solving classification problems.

KNN algorithms were proposed by T. M. Cover and P. E. Hart in 1967. The algorithm is used by making use of the data in a sample set with certain classes. The distance of the new data, which will be added to the sample data set, is calculated according to the existing data, and its k close neighbors are checked. Three types of distance functions are generally used for distance calculations: Euclidean Distance, Manhattan Distance or Minkowski Distance. (Pandey, A., & Jain, A., 2017).

KNN; It is one of the most popular machine learning algorithms due to its resistance to old, simple and noisy training data. But it also has a disadvantage. For example, it needs a lot of

memory space when used for large data, since it stores all states when calculating distances.

The steps of the KNN algorithm:

- First, the parameter k is determined. This parameter is the number of nearest neighbors to a given point. For example: Let k=2. In this case, classification will be made according to the 2 closest neighbors.
- The distance of the new data to be included in the sample data set is calculated one by one according to the existing data with the help of the corresponding distance functions.
- The k nearest neighbors of the related distances are considered. It is assigned to the class of k neighbors or neighbors according to the attribute values.
- The selected class is considered to be the class of the observation value expected to be estimated. In other words, the new data is labeled.

2.2.2. Random Forest Algoritm (RF)

Random forest algorithm is one of the supervised classification algorithms. It is used in both regression and classification problems. The algorithm aims to increase the classification value during the classification process by producing more than one decision tree. Random forest algorithm is the process of choosing the highest score among many decision trees that work independently of each other. As the number of trees increases, our rate of obtaining a precise result increases. The main difference between the decision tree algorithm and the random forest algorithm is that the process of finding the root node and splitting the nodes is random.

2.2.3. Extreme Gradient Boosting Algoritm (XGB)

XGBoost(eXtreme Gradient Boosting) is a high-performance version of the Gradient Boosting algorithm optimized with various modifications. It was developed by Tianqi Chen and Carlos Guestrin in 2016. (Chen, T., & Guestrin, C., 2016) The most important features of the algorithm are that it can achieve high predictive power, prevent over-learning, manage empty data and do them quickly. According to Tianqi, XGBoost runs 10 times faster than other popular algorithms.

Software and hardware optimization techniques have been applied to obtain superior results using less resources. It is cited as the best of the decision tree-based algorithms.

Instead of examining each value in the data, XGBoost divides the data into pieces (quantiles) and works according to these pieces. As the amount of parts is increased, the algorithm will look at smaller intervals and make better predictions.

2.3. Evaluation Metrics

Evaluation metrics are used to measure the quality of the model. Evaluating the success of the developed models or projects is important for both the realized and future works. To evaluate the performance of classification models used in machine learning, the confusion matrix, which compares the predictions of the target attribute and the actual values, is often used. Confusion Matrix is given Table 3. (Zeng, G., 2020)

Table 3. Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Based on the confusion matrix, accuracy, recall, precision and F1-score parameters are also reached.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1 - Score = 2\frac{1}{\frac{1}{\frac{TP}{TP + FP}} + \frac{1}{\frac{TP}{TP + FN}}} \quad (4)$$

These evaluation metrics in 1,2,3,4 can be used together to determine to success of each model. (Dalianis, H., 2018)

3. Data Collection, Pre-Processing and Visualization of Crime Dataset

The Crime dataset consists of twenty-three columns and 1.4 million rows and was downloaded from the open-source dataset kaggle site[†]. Also, Chicago's 2012-2016 weather records[‡] and Chicago Community Area population information from wikipedia were obtained. In this study, unlike other crime studies, the effect of population and weather on crime was investigated by creating a data set that combines three separate sources. In the Data Collection section, the features that will be evaluated in the study to be done by removing the columns that have very few records from the combination of the three data sets or that are unique such as id and that will not affect the study are selected and their details are given in Table 4. In the Data Pre-processing section, it is mentioned how new features are created based on existing data and which transformation operations are performed on the data. The last version of the data set is shown in Table 5. In the Data Visualization section, graphical evaluations are given over the data set arranged. As a result of all operations, it is aimed to influence the sampling techniques to make predictions on Arrest, Crime Type and Security fields.

3.1. Data Collection

The data set used in the study consists of a combination of data from three diverse sources. The data consists of five years of criminal records from the Chicago Police Department. The weather information of the date of each criminal record has been added to the data set based on date. Finally, information such as population and size of the regions by meter square where the crime was committed was added to the data set, allowing the data set to be examined with expanded different perspectives. In Table

[†] https://www.kaggle.com/datasets/currie32/crimes-in-chicago?select=Chicago_Crimes_2012_to_2017.csv

[‡] https://www.ncdc.noaa.gov/

4, attribute information of the combined data set used in the analysis is given.

Table 4. Attributes of Combined Dataset

Column Name	Description	Type	Number of Distinct Value	Min Value	Max Value
Date	Date when the incident occurred.	date	1827	01-01-12	31-12-16
Primary Type	The primary description of the IUCR code.	object	32	-	-
Location Description	Description of the location where the incident occurred.	object	145	-	-
Arrest	Indicates whether an arrest was made.	bool	2	-	-
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	bool	2	-	-
Community Area	Indicates the community area where the incident occurred.	integer	77	1	77
Name	Community Area Name	object	77	-	-
max_tmp	The maximum temperature for that day in fahrenheit	integer	100	-2	103
min_tmp	The minimum temperature for that day in fahrenheit	integer	95	-18	82
Population	The population of Comunity Area of Chicago	integer	77	2527	105481
Area_km2	The value of the area in square kilometers	integer	69	1.5	34.55
Density_km2	The density of the area in square kilometers	integer	77	388.36	14863.58

3.2. Class Distribution of Arrest, Crime Type and Security Target Area

An imbalanced dataset is the result of an imbalance between the data of the class in the target field. In general, one class is much more or less than the other class. Working on the data of imbalanced classes poses a problem in terms of the reliability of the results.

In both the train data and test data obtained from the unbalanced dataset, the majority belong to the class. Therefore, the data is more prone to yield high results. But this is related to the fact that the data is unbalanced. It is more difficult to obtain high results by using a balanced data set where each class is equal. Because the train and test data are selected from the data set belonging to the same class, and the difference between the classes is very small. The success obtained from balanced data also means that the success obtained from unbalanced data is much more consistent and the model is suitable for the data set.

Sampling techniques were applied for three different target areas in order to eliminate the problem in imbalanced data. Figure 1, Figure 2 and Figure 3 show the distribution of classes for each target area.

3.2.1. Target Arrest

There are two classes for the Arrest target area, True and False. There are a total of %73,5 False and %26,5 True records in all data.

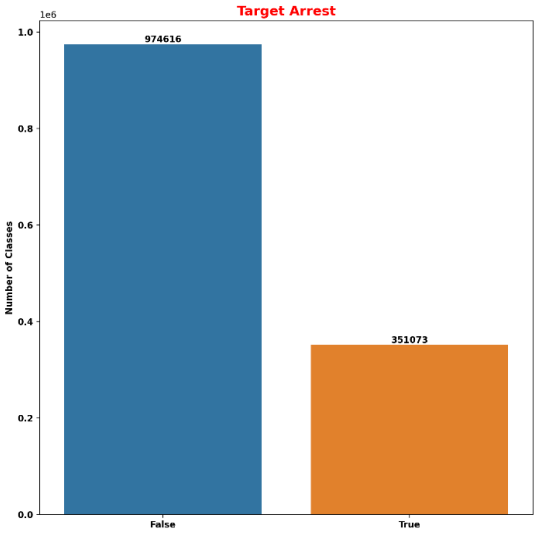

Figure 1. Classes of Arrest

3.2.2. Target Crime Type

In all data, the Crime Type target field consists of three classes: %44,25 materially damaging crime, %39,6 physically damaging crime and %16,15 emotionally damaging crime.

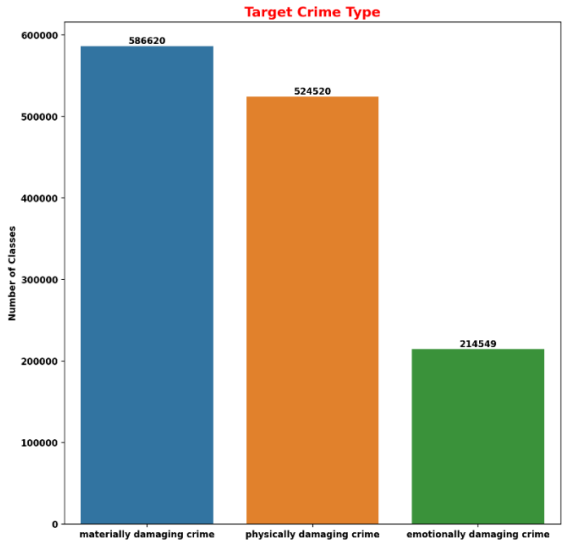

Figure 2. Classes of Crime Type

3.2.3. Target Security

There are four different classes for the Security target area. Two of them are close to each other while the other two are far from each other. Out of the whole data set, %50,6 of them belong to the very dangerous class, %23 of them to the safe class, %20,4 to the dangerous class and %6 to the very safe class.

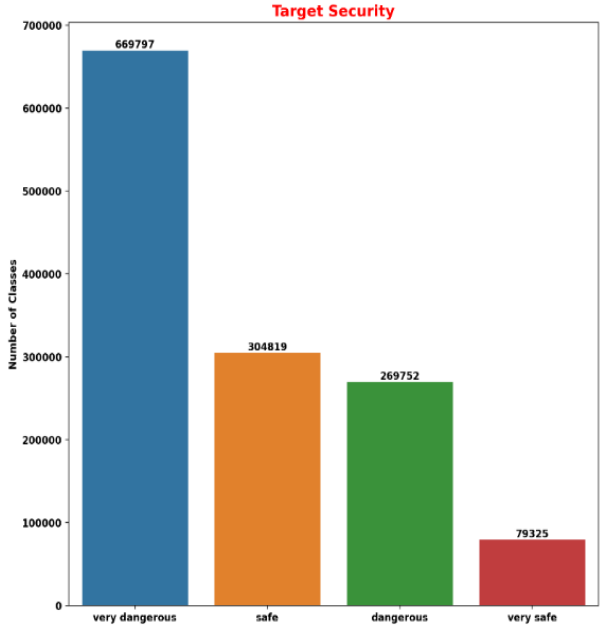

Figure 3. Classes of Security

3.3. Data Visualization of Crime Dataset

In this section, the visualization of the data and the effect of the attributes on the data through these graphs are mentioned.

3.3.1. Location, Population and Crime Relationship

To analyze the relationship between population, location and crime, the population and crime amounts corresponding to each location were examined and visualized.

In Figure 4, the population information of each community area is given from the highest population to the lowest population.

In Figure 5, the average crime amount is given from the highest amount of crime to the lowest amount of crime in each community area. Thus, the community area information above and below the crime average is also included.

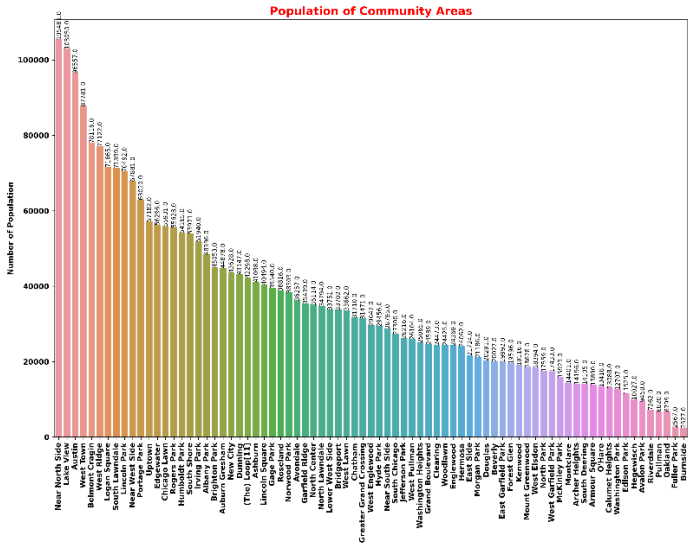

Figure 4. Population of Comunity Areas

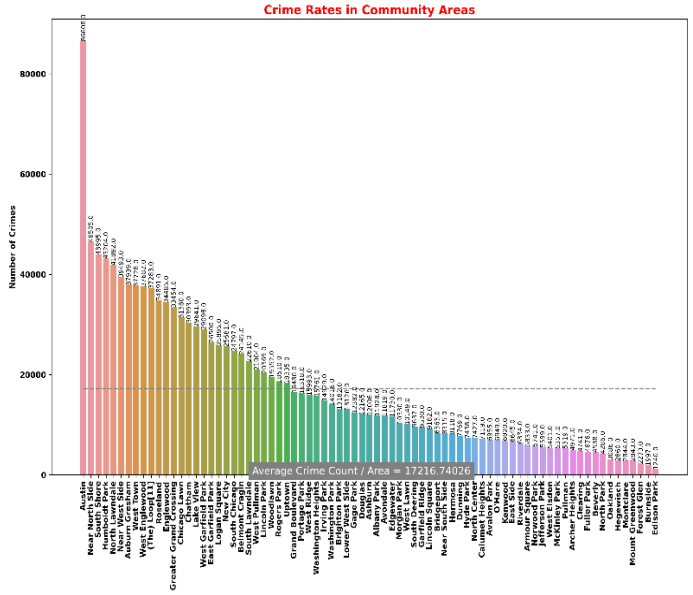

Figure 5. Average Crime Count of Comunity Areas

The highest crime rate was seen in Austin, Chicago's third most populous comunty area, by a wide margin. The Near North Side, with the highest population, is the second most crime-ridden area. Although the South Shore is not a very crowded area, it has been observed that it is the third region where crime is committed the most. It has been observed that the crime rate is low in places where the population is low, such as Burnside and Edison Park.

Based on all this, although the population and crime rate are relatively proportional, regions such as Austin and South Shore have been the regions where crime is high regardless of the population.

3.3.2. Change of All Crimes over the Years

Figure 6 contains detailed information about each criminal record. The total number of crimes committed for each year and its change over the years are shown in the graphs.

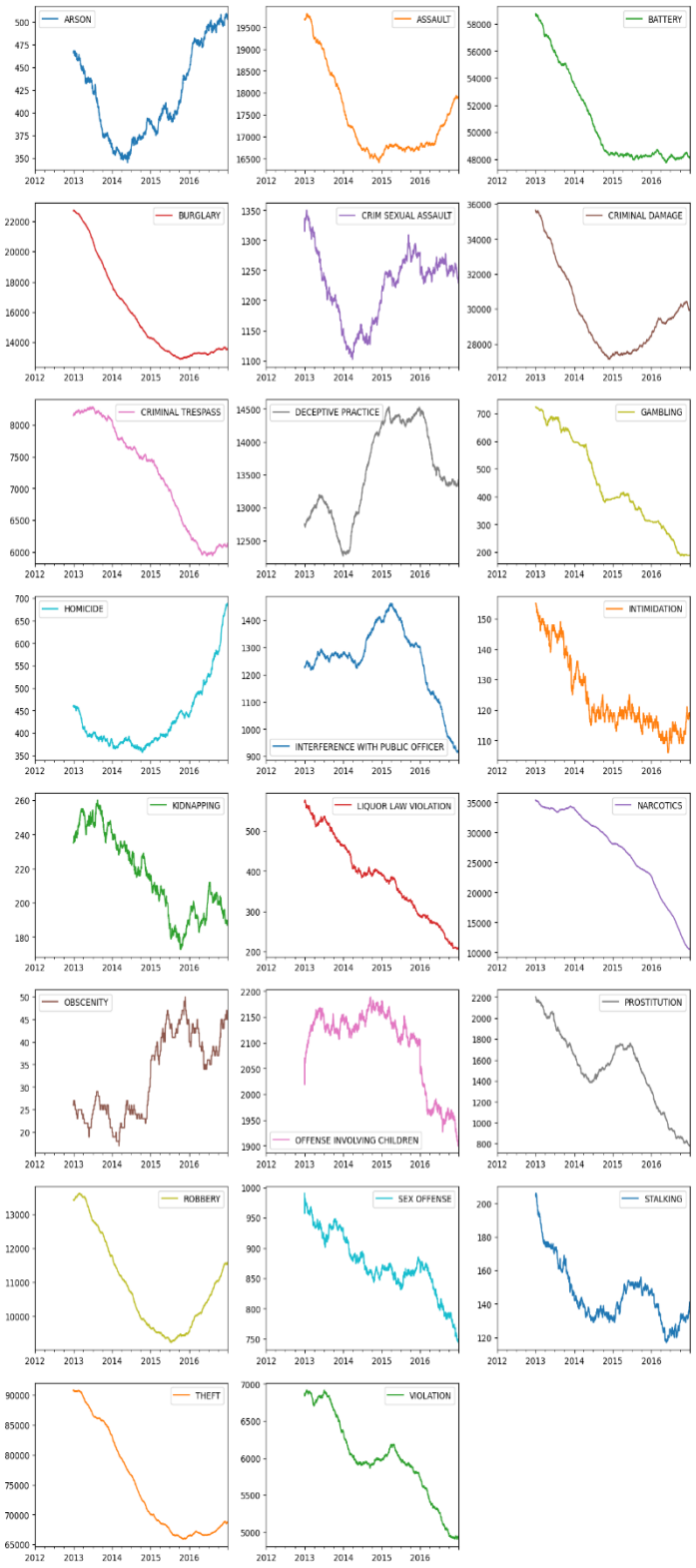

Figure 6. Change of All Crimes

The following inferences were obtained from the crime graphs:

- Arson and Homicide is the most increasing type of crime over the years,
- Gambling, Narcotics, Violation, Prostitution, Interference with Public Officer, Liquor Law Violation and Offense Involving Children crime types are almost not seen or seen truly little in recent years,
- Battery, Criminal Trespass and Burglary crime types have less momentum in recent years,
- The ups and downs in crimes are mostly seen in Crim Sexual Assault, Deceptive Practice and Obsencity,
- In general, it has been determined that all crimes have fluctuations in the graphics, some years more and some years less.

3.3.3. Season, Weather and Crime Relationship

Twelve months in our dataset were divided into four seasons to facilitate review. The first three months constitute the first season, and each subsequent three months constitute the next season. In Figure 7, each of the four episodes between each year corresponds to a season.

Chicago is located in North America. At the same time, it has a climate where the summer months are hot in the middle of the year and the winter months are cold at the end of the year.

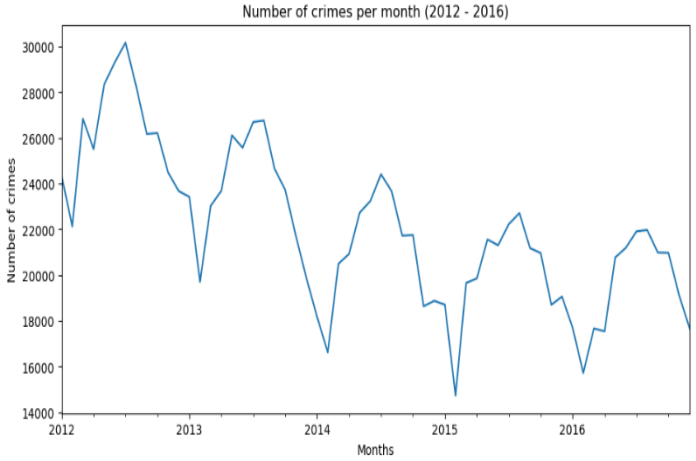

Figure 7. Change in Crime Rates over the Years

In general, the highest crime rates were determined in 2012, and the lowest crime rates were determined in 2016. It has been observed that crime rates always decrease at the beginning of each year, and reach their highest levels in the middle of the year. From this point of view, it is seen that the most crimes are committed in the summer months when the weather is hot, and the least crimes are committed in the winter months when the weather is cold. Thus, it is observed that there is a direct proportional relationship between temperature and seasons and crime.

3.4. Data Pre-Processing

A lot of data pre-processing steps are applied in the period from the first acquisition of the data to the implementation of the algorithms. In this section, the details of these processes are mentioned.

3.4.1. Handling Missing Data and Extract Noisy Data

Data deletion method was used for missing values. Similarly, information with values far below normal was removed so that it would not destabilize the data set. For example, there was twenty value whose Primary Type record was "HUMAN TRAFFICKING". This was around 0% when compared to other types of crime. Therefore, all data containing this record was deleted from the data set. The same procedure was repeated for similar samples. Noisy data was removed from the data set so that it would not adversely affect the success of the model.

3.4.2. Data Integration and Transformation

In the processing of raw data, the data set undergoes various transformations. This is sometimes the process of removing unnecessary data. Sometimes, it may produce a new feature from the data, especially in machine learning projects.

- Mean_temp_C: The maximum and minimum temperature values kept in fahrenheit in the data were first converted to celsius. Later, a new feature was added by averaging the two values.
- Season: The month information was removed from the date information. According to the month information, season information as spring, summer, autumn and winter has been added.
- Weekday Weekend: By using the python function from the date information, the information of whether the day of the relevant date is weekend or weekday was added.
- When: Two-class time information as morning or evening has been added from the AM and PM information in the date information.
- Security: While creating the Security area, 2 points were taken as a basis. First of all, the number of crimes committed in all community areas, their percentage and the average amount of crime were found. Based on this information, 4 different levels of security were assigned as very safe, safe, dangerous, and very dangerous by looking at the number of crimes committed in the community area and evaluating the relationship with whether the suspect was caught or not.
    - It is very dangerous if more than average crime has been committed and the suspect has not been arrest,
    - Dangerous if more than average crime has been committed and the suspect has been arrested,
    - Safe if a below-average crime has been committed and the suspect has not been arrested,
    - Considered very safe if below-average crime was committed and the suspect was arrested.
- Crime Type: According to the crime information in the Primary Type field, crime types were determined as physically, materially, emotionally damaging crime. For example, “Theft” was determined as material, “Battery” as physical, “sex Offense” as emotionally damaging crime.

3.4.3. Encoding Method

One-hot Encoding technique was applied for all categorical values. According to this technique, a new column is opened for each class of categorical value. If the value in the data contains that class, one is assigned, if it does not, zero is assigned and a dummy value is given. With this method, all categorical values are converted to numerical values.

Table 5. AttributesTthat Applied Encoding Method

Attribute	Data Type	Encoding Format
Primary Type	categorical	One-Hot Encoding
Location Description	categorical	One-Hot Encoding
Arrest	categorical	One-Hot Encoding
Domestic	categorical	One-Hot Encoding
Name	categorical	One-Hot Encoding
season	categorical	One-Hot Encoding
Weekday Weekend	categorical	One-Hot Encoding
when	categorical	One-Hot Encoding
security	categorical	One-Hot Encoding
Crime Type	categorical	One-Hot Encoding

3.4.4. Normalization

Min-max normalization was applied for each numerical value. Thus, the range in the numerical values in the dataset has been narrowed and all the data has been moved to the 0-1 range. This method was also applied to prevent uneven distribution in the dataset.

Table 6. Attributes That Applied Normalization Method

Attribute	Data Type	Normalization Type
Population	numerical	Min-Max Normalization
Area_km2	numerical	Min-Max Normalization
Density_km2	numerical	Min-Max Normalization
Mean_temp_C	numerical	Min-Max Normalization

3.4.5. Description of Pre-Processed Dataset

The data set was formed by combining the crime data recorded by the Chicago Police Department between 2012 and 2016 with the weather and population data by date. In total, when all the pre-processing of the data is completed, there are 18 columns 1323693 records given to the algorithms.

Table 7. Attributes of Pre-processed Dataset

Column Name	Description	Type	Method	Related Columns	Values	Value Range
Primary Type	The primary description of the IUCR code.	int	encoding	23	0,1	0-1
Location Description	Description of the location where the incident occurred.	int	normalization	29	0,1	0-1
Domestic	Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.	int	encoding	2	0,1	0-1
Arrest	Indicates whether an arrest was made.	int	encoding	2	0,1	0-1
Name	Community Area Name	int	encoding	77	0,1	0-1
Population	The population of Comunity Area of Chicago	int	normalization	1	0.0,…,0.99	0-1
Area km2	The value of the area in square kilometers	int	normalization	1	0.0,…,0.99	0-1
Density km2	The density of the area in square kilometers	int	normalization	1	0.0,…,1.0	0-1
Mean	The average temperature for that day in celsius	int	normalization	1	0.0,…,0.99	0-1

temp_C						
Column Name	**Description**	**Type**	**Method**	**Related Columns**	**Values**	**Value Range**
season	Season information according to the time the crime was committed	int	encoding	4	0,1	0-1
Weekday Weekend	Weekday or weekend information according to the time the crime was committed	int	encoding	2	0,1	0-1
when	AM (morning) or PM (evening) information according to the time the crime was committed	int	encoding	2	0,1	0-1
security	Security information evaluated according to the average crime rate of the area where the crime was committed and whether the suspect was arrest or not.	int	encoding	4	0,1	0-1
Crime Type	Information on how the crime committed according to the Primary Type information causes harm.	int	encoding	3	0,1	0-1

4. Experiments Setup

The data consists of a combination of 3 sources. Combining these data described in the Data Collectin section was performed by writing python codes in a Jupyter notebook and using the pandas library. Processes such as examining the data, removing missing data, and performing the necessary numerical operations were made with the use of pandas and numpy libraries.

All graphics in the data visualization section are made using python's data visualization libraries, scipy, matplotlib, seaborn. By visualizing the data, it was examined in detail what kind of data was worked on before the algorithms.

In the data pre-processing part, cleaning of missing data in the dataset, data transformation, new feature formation, encoding and normalization processes were performed. All operations are done using python. Sklearn was used for normalization operations, pandas and numpy library were used for all other operations.

After all data pre-processing, the final dataset to be used in sampling algorithms has been reached.

Similar libraries were also used with the Data visualization step in the use of target areas as graphics.

However, as a result of these operations, the size of the data has grown enormously. Principal Component Analysis (PCA) (Wold, S., Esbensen, K., & Geladi, P., 1987) and Linear Discriminant Analysis (LDA) (Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., & Erni, F., 1996) are commonly used methods for dimensionality reduction in the literature. In this study, PCA method was preferred.

It is a technique whose main purpose is to keep the data set with the highest variance in high-dimensional data, but to provide dimension reduction while doing this. By finding the general features in the over-dimensional data, it reduces the number of dimensions and compresses the data. It is certain that some features will be lost with size reduction; but the intent is that these disappearing traits contain little information about the population. This method combines highly correlated variables to create a smaller set of artificial variables, called "principal components", that make up the most variation in the data. In this article, the data is restated with 2 components and shown in Figure 8, Figure 9 and Figure 10 in two dimensions according to the targets.

PCA generally consists of five steps:

1. In the first step, the data is centered by subtracting the average from each variable. Thus, a data set with a mean of 0 is obtained. However, if the variances of the original dataset are very different from each other, the data can be converted between 0 and 1.
2. Secondly, the covariance and correlation matrix is created. In scaled data, both are the same.
3. Then the eigenvalues and eigenvectors of the covariance and correlation matrix are calculated.
4. The eigenvalues are ordered from largest to smallest and the corresponding eigenvectors are found. Thus, the principal components are selected.
5. Finally, after all these processes, a new data set is created.

PCA is also used to visualize data along with dimensional reduction. Data reduced to two dimensions with PCA are shown in figure 8 for the arrest target area. Red color indicates true class and blue color indicates false class.

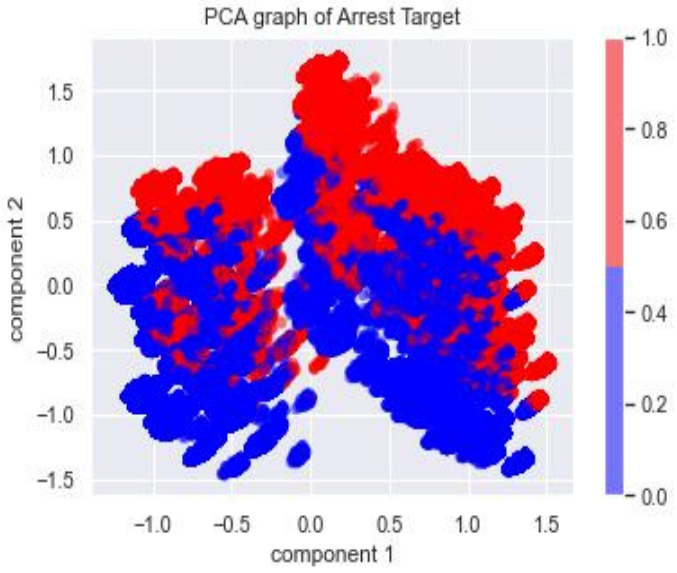

Figure 8. Graphical Representation of the Arrest Target After Applying PCA

In Figure 9, red color represents emotionally damaging crime class, blue color represents physically damaging crime class and

yellow color represents materially damaging crime class for crime type target.

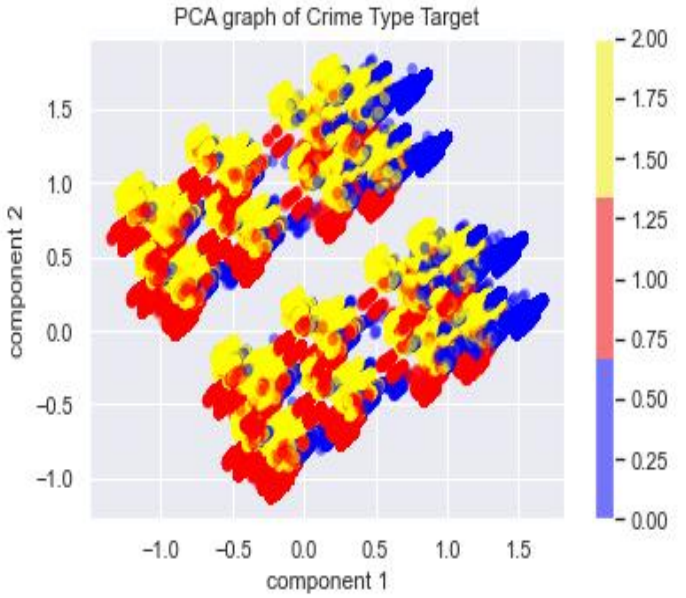

Figure 9. Graphical Representation of the Crime Type Target After Applying PCA

In Figure 10, blue color refers very safe class, red color refers safe class and yellow color refers dangerous class and white color refers very dangerous class for security target.

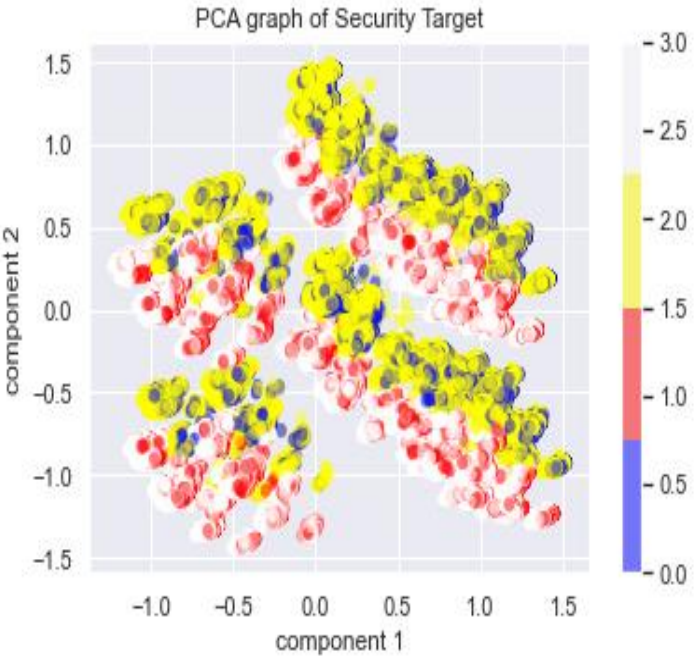

Figure 10. Graphical Representation of the Crime Type Target After Applying PCA

While applying the PCA algorithm, the sklearn library was used. Matplotlib and seaborn libraries were used for data visualization with PCA.

By using machine learning methods, models suitable for large data sets are obtained. Evaluation is helpful to find out which model is better and to understand how well the learning model will work in the future. Using only the training set is not acceptable because the method may be overfitting. Two different approaches are generally recommended to avoid this situation: hold-out and cross-validation. Both approaches use a test set to avoid overfitting and measure model performance. In this study, the hold-out method was used.

Hold-out is a method of dividing the dataset into training and test sets. The training set is the data on which the model is trained, while the test set is the data used to see how well the model performs on the untrained data. Commonly when using the hold-out method, it uses 80% of the data for training and 20% of the remaining data for testing.

The hold-out method is especially useful for very large datasets. Because cross validation uses more than one training-test split, it takes more time than hold-out because it requires more computational power.

The data was divided into two using the hold-out method. 80% of it was considered as training data and 20% as test data.For classification studies, 3 different targets were determined as Arrest, Security and Crime Type.

The sklearn library was used when applying the hold-out method and calculating the evaluation metrics such as accuracy, precision, recall, F1-score.

Sampling algorithms, which were evaluated comparatively from four different perspectives (Statistic: Stratify Sampling, Over-Sampling: Random Over Sampling, Synthetic Minority Over-sampling, Adaptive Synthetic Sampling, Under-Sampling: Random Under Sampling, Near Miss Sampling, Neighborhood Cleaning Rule, Mix-Sampling: Smote Tomek) were applied for the target areas of arrest, crime type and security by using three different classification algorithms (K-Nearest Neighborhood, Random Forest, Extreme Gradient Boosting).

For each target area, a total of twenty-four tests were conducted using eight sampling and three classification algorithms, and a total of seventy-two tests for all target areas. Moreover, for three different target areas and three different classification algorithms, a total of nine tests were conducted without using the sampling method, and a total of eighty-one tests were conducted as a result of all studies.

5. Computational Result and Future Work

All training results were analyzed in detail and presented in three tables. Each sample is listed in a way that is comparable to both other sampling methods and the sampling methods in its group.

The results of all experiments for Arrest Target in Table 8, Crime Type Target in Table 9 and Security Target in Table 10 are given. High success rate sampling techniques and classification algorithms are marked in bold. Evaluations related to each table are given below the table, and the general evaluation is given at the end.

When all the results were evaluated in general, high success rates were determined for all three target areas. It has been observed that with the use of sampling techniques, the imbalanced data set becomes more stable and the success rates increase for each target area.

Since this study was based on crime data from the Chicago Police department, the coverage was limited to one city. More comprehensive and general results will be obtained if detailed crime analysis and machine learning studies are carried out with

the crime records in the police departments or FBI databases across the country.

Table 8. The Result of Target Arrest

Target: Arrest			Pre-Processed Data			
			Accuracy	**Precision**	**Recall**	**F1 Score**
Imbalance Dataset		KNN	0.98	0.98	0.98	0.98
		RF	0.98	0.98	0.98	0.98
		XGB	0.96	0.96	0.96	0.96
Statistic	SS	KNN	0.98	0.98	0.97	0.97
		RF	0.98	0.98	0.97	0.97
		XGB	0.96	0.96	0.93	0.94
Over Sampling	**ROS**	**KNN**	**0.99**	**0.99**	**0.99**	**0.99**
		RF	**0.99**	**0.99**	**0.99**	**0.99**
		XGB	0.96	0.96	0.93	0.94
	SMOTE	KNN	0.98	0.98	0.98	0.98
		RF	0.98	0.98	0.98	0.98
		XGB	0.95	0.95	0.94	0.94
	ADASYN	KNN	0.97	0.97	0.97	0.97
		RF	0.97	0.97	0.97	0.97
		XGB	0.87	0.87	0.87	0.87
Under Sampling	RUS	KNN	0.97	0.97	0.97	0.97
		RF	0.98	0.98	0.97	0.97
		XGB	0.96	0.96	0.93	0.94
	NMS	KNN	0.96	0.96	0.96	0.96
		RF	0.96	0.96	0.96	0.96
		XGB	0.92	0.93	0.92	0.92
	NCR	**KNN**	**0.99**	**0.99**	**0.99**	**0.99**
		RF	**0.99**	**0.99**	**0.99**	**0.99**
		XGB	**0.96**	**0.96**	**0.96**	**0.96**
Mix (Both Over and Under SAmpling)	**ST**	**KNN**	**0.99**	**0.99**	**0.99**	**0.99**
		RF	**0.99**	**0.99**	**0.99**	**0.99**
		XGB	0.95	0.95	0.95	0.95

According to Table 8, it has been determined that the most compatible sampling method for the arrest target area is ROS, NCR and ST, which are used together with KNN and RF. Even the XGB algorithm is used, it has been seen that the best sampling method is NCR. It has been observed that NCR, which works in harmony with all classification algorithms, achieves the best success among all sampling methods. Although the accuracy rate of the imbalance data is high, it has been determined that the

sampling methods both make the accuracy of the imbalance data more consistent and increase it by 1%.

Table 9. The Result of Target Crime Type

Target: Crime Type			Pre-Processed Data			
			Accuracy	Precision	Recall	F1 Score
Imbalance Dataset		KNN	0.94	0.94	0.94	0.94
		RF	0.94	0.94	0.94	0.94
		XGB	0.85	0.86	0.85	0.86
Statistic	SS	KNN	0.98	0.98	0.97	0.97
		RF	0.94	0.93	0.93	0.93
		XGB	0.86	0.86	0.86	0.86
Over Sampling	ROS	KNN	0.94	0.93	0.94	0.93
		RF	0.94	0.93	0.93	0.93
		XGB	0.86	0.86	0.86	0.86
	SMOTE	KNN	0.95	0.95	0.95	0.95
		RF	0.95	0.95	0.95	0.95
		XGB	0.86	0.87	0.86	0.86
	ADASYN	KNN	0.93	0.93	0.93	0.93
		RF	0.93	0.93	0.93	0.93
		XGB	0.80	0.80	0.80	0.80
Under Sampling	RUS	KNN	0.94	0.94	0.94	0.94
		RF	0.94	0.94	0.94	0.94
		XGB	0.88	0.88	0.88	0.87
	NMS	KNN	0.89	0.89	0.89	0.89
		RF	0.89	0.89	0.89	0.89
		XGB	0.81	0.81	0.81	0.81
	NCR	**KNN**	**0.98**	**0.98**	**0.98**	**0.98**
		RF	**0.98**	**0.98**	**0.98**	**0.98**
		XGB	**0.88**	**0.89**	**0.88**	**0.88**
Mix	ST	KNN	0.96	0.96	0.96	0.96
		RF	0.96	0.96	0.96	0.96
		XGB	0.87	0.88	0.87	0.87

According to Table 9, the most successful sampling method for the crime type target area was NCR. It has also been observed that it is compatible with all classification algorithms. The ST sampling method, which is a combination of over and under sampling, has also been found to be the second most successful

sampling method after NCR. At the same time, the success rate was increased by 4% with these sampling methods.

Table 10. The Result of Target Security

Target: Security			Pre-Processed Data			
			Accuracy	Precision	Recall	F1 Score
Imbalance Dataset		KNN	0.90	0.90	0.90	0.90
		RF	0.89	0.89	0.89	0.89
		XGB	0.74	0.74	0.74	0.69
Statistic	SS	KNN	0.90	0.90	0.90	0.90
		RF	0.89	0.89	0.89	0.89
		XGB	0.74	0.74	0.74	0.69
Over Sampling	ROS	KNN	0.94	0.94	0.94	0.94
		RF	0.94	0.93	0.93	0.93
		XGB	0.75	0.76	0.75	0.69
	SMOTE	KNN	0.94	0.94	0.94	0.94
		RF	0.93	0.93	0.93	0.93
		XGB	0.69	0.70	0.69	0.69
	ADASYN	KNN	0.91	0.91	0.91	0.91
		RF	0.90	0.90	0.90	0.90
		XGB	0.62	0.62	0.62	0.61
Under Sampling	RUS	KNN	0.90	0.89	0.90	0.89
		RF	0.89	0.88	0.89	0.88
		XGB	**0.80**	**0.81**	**0.80**	**0.74**
	NMS	KNN	0.89	0.89	0.89	0.89
		RF	0.88	0.88	0.88	0.88
		XGB	0.78	0.78	0.78	0.78
	NCR	KNN	0.96	0.96	0.96	0.96
		RF	**0.95**	**0.95**	**0.95**	**0.95**
		XGB	0.76	0.77	0.76	0.70
Mix	**ST**	**KNN**	**0.98**	**0.95**	**0.95**	**0.95**
		RF	**0.95**	**0.95**	**0.95**	**0.95**
		XGB	0.69	0.70	0.69	0.69

According to Table 10, in general, it was seen that the most successful sampling method for the security target area was ST. However, it has been determined that the RUS sampling method and the XGB algorithm perform better than the ST sampling

method together with the XGB algorithm. Sampling methods increased the success rate the most with the field of Security by 8% compared to the imbalanced data.

Increasing the success rate with the sampling method is more difficult as all classes are equal and there is no tendency to predict any class better. However, we were able to increase the success rate with sampling methods when our goal was to have two classes, three classes or four classes. We also found that the success rate of balance dataset increases more than imbalance dataset when the number of classes is large.

In addition to this, studies were carried out on eight sampling methods. In the next step, a more detailed study can be done by expanding the scope of the study by adding all sampling methods, few classification algorithms and a few more validation methods.

References

Hibberts, M., Burke Johnson, R., & Hudson, K. (2012). Common survey sampling techniques. In *Handbook of survey methodology for the social sciences* (pp. 53-74). Springer, New York, NY.

Zhihao, P., Fenglong, Y., & Xucheng, L. (2019, April). Comparison of the different sampling techniques for imbalanced classification problems in machine learning. In 2019 11th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA) (pp. 431-434). IEEE.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. ACM SIGKDD explorations newsletter, 6(1), 20-29.

Sathyadevan, S., Devan, M. S., & Gangadharan, S. S. (2014, August). Crime analysis and prediction using data mining. In 2014 First international conference on networks & soft computing (ICNSC2014) (pp. 406-412). IEEE.

Junsomboon, N., & Phienthrakul, T. (2017, February). Combining over-sampling and under-sampling techniques for imbalance dataset. In Proceedings of the 9th International Conference on Machine Learning and Computing (pp. 243-247).

Prabakaran, S., & Mitra, S. (2018, April). Survey of analysis of crime detection techniques using data mining and machine learning. In Journal of Physics: Conference Series (Vol. 1000, No. 1, p. 012046). IOP Publishing.

Xie, C., Du, R., Ho, J. W., Pang, H. H., Chiu, K. W., Lee, E. Y., & Vardhanabhuti, V. (2020). Effect of machine learning re-sampling techniques for imbalanced datasets in 18F-FDG PET-based radiomics model on prognostication performance in cohorts of head and neck cancer patients. European journal of nuclear medicine and molecular imaging, 47(12), 2826-2835.

Etikan, I., & Bala, K. (2017). Sampling and sampling methods. Biometrics & Biostatistics International Journal, 5(6), 00149.

Albahli, S., Alsaqabi, A., Aldhubayi, F., Rauf, H. T., Arif, M., & Mohammed, M. A. (2021). Predicting the type of crime: Intelligence gathering and crime analysis. Computers, Materials & Continua, 66(3), 2317-2341.

Kurin, S., Steinshamn, S. I., & Saerens, M. (2017). " A comparison of classification models for imbalanced datasets.

Meng, X. (2013, May). Scalable simple random sampling and stratified sampling. In International Conference on Machine Learning (pp. 531-539). PMLR.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

Mani, I., & Zhang, I. (2003, August). kNN approach to unbalanced data distributions: a case study involving information extraction. In Proceedings of workshop on learning from imbalanced datasets (Vol. 126, pp. 1-7). ICML.

DURAHİM, A. O. (2016). Comparison of sampling techniques for imbalanced learning. Yönetim Bilişim Sistemleri Dergisi, 2(2), 181-191.

Pandey, A., & Jain, A. (2017). Comparative analysis of KNN algorithm using various normalization techniques. International Journal of Computer Network and Information Security, 9(11), 36.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Zeng, G. (2020). On the confusion matrix in credit scoring and its analytical properties. Communications in Statistics-Theory and Methods, 49(9), 2080-2093.

Dalianis, H. (2018). Evaluation metrics and evaluation. In Clinical text mining (pp. 45-53). Springer, Cham.

Laurikkala, J. (2001, July). Improving identification of difficult small classes by balancing class distribution. In Conference on artificial intelligence in medicine in Europe (pp. 63-66). Springer, Berlin, Heidelberg.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008, June). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence) (pp. 1322-1328). IEEE.

Tomek, I. (1976). Two modifications of CNN. IEEE Trans. Systems, Man and Cybernetics, 6, 769-772.

Batista, G. E., Bazzan, A. L., & Monard, M. C. (2003, December). Balancing Training Data for Automated Annotation of Keywords: a Case Study. In WOB (pp. 10-18).

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.

Wu, W., Mallet, Y., Walczak, B., Penninckx, W., Massart, D. L., Heuerding, S., & Erni, F. (1996). Comparison of regularized discriminant analysis linear discriminant analysis and quadratic discriminant analysis applied to NIR data. Analytica Chimica Acta, 329(3), 257-265.