Çukurova Üniversitesi Mühendislik Fakültesi Dergisi, 38(4), ss. 1035-1048, Aralık 2023
Cukurova University Journal of the Faculty of Engineering, 38(4), pp. 1035-1048, December 2023

A Design of Crime Category Detection Framework using Stacking Ensemble Model

Recep Sinan ARSLAN*[1] ORCID *0000-0002-3028-0416*
Burak DÜLGEROĞLU[2] ORCID *0009-0000-8201-9343*

[1]Kayseri University, Faculty of Engineering, Architecture and Design, Department of Computer Engineering, Kayseri, Türkiye
[2]Kayseri University, Graduate School of Education, Kayseri, Türkiye

Geliş tarihi: 16.10.2023 Kabul tarihi: 25.12.2023

Atıf şekli/ How to cite: ARSLAN, R.S., DÜLGEROĞLU, B., (2023). A Design of Crime Category Detection Framework using Stacking Ensemble Model. Cukurova University, Journal of the Faculty of Engineering, 38(4), 1035-1048.

Abstract

Crime refers to an action legally defined as harmful to society, and it is important to understand the type of crime to prevent these actions. However, crime can occur at any time and place, making it difficult to predict. Data generated based on previously committed crimes contributes to overcoming this difficulty. This study proposes a novel model for classifying criminal activities using a Doc2Vec that can cause a numerical representation of texts regardless of length and a stacking ensemble model that includes 8 different machine-learning models. Unlike the literature, the model processes the features as text and converts them into vectors rather than categorically. In this way, it enables using features that cannot be used in the literature. The proposed model is tested using a distributed online competition database, Francisco Crime Classification, which contains crimes committed over 12 years. An accuracy value of 99.28% was obtained for the 15 crime categories with the highest crime records, while precision, recall, and f-score values were 99.18%, 99.38%, and 99.20%, respectively. With cross-validation (k=10), 99.80% performance was achieved with a std. value of 0.001. These performance values are higher than those of all the studies in the literature using categorical feature structures. The results show that converting criminal activity reports, which contain text-based features, into vectors that can be processed with natural language processing techniques such as Doc2vec instead of using them categorically in model training can directly contribute to the classification performance and provide a more efficient model with less preprocessing.

Keywords: Crime prediction, Criminology, Doc2vec, Stacking ensemble model

Suç Kategorisi Tespiti için Yığınlama Topluluk Öğrenimi Modeli Kullanan Çatı Tasarımı

Öz

Suç, toplum açısından kanuni olarak zararlı olarak tanımlanmış eylemi ifade eder ve bu eylemlerin engellenmesi için suç türünün anlaşılması oldukça önemlidir. Ancak suç herhangi bir zamanda ve yerde

[*] Sorumlu yazar (Corresponding Author): Recep Sinan ARSLAN, *sinanarslanemail@gmail.com*

meydana gelebilmektedir ve bu durum suçun tahmin edilebilirliğini zorlaştırmaktadır. Daha önce işlenmiş suçlara dayalı olarak oluşturulan verilerin kullanılması bu zorluğun aşılmasına katkı sağlamaktadır. Bu çalışmada suç faaliyetlerini sınıflandırma için uzunluğundan bağımsız olarak metinlerin sayısal temsilini üretebilen Doc2Vec yapısı ve 8 farklı yapay öğrenme modelini içeren bir yığınlama topluluk öğrenimi modelin kullanıldığı özgün bir model önerilmiştir. Model literatürden farklı olarak öznitelikleri kategorik olarak değil metin olarak işlemekte ve vektör haline dönüştürmektedir. Bu sayede literatürde kullanılamayan özniteliklerin kullanılmasını sağlamaktadır. Önerilen model 12 yıl boyunca işlenen suçları içeren, Francisco Crime Classification, isimli online dağıtımlı bir çevrimiçi yarışma veriseti kullanılarak test edilmiştir. En yüksek suç kaydının olduğu 15 suç kategorisi için %99,28 doğruluk değeri elde edilirken, kesinlik, geri çağırma ve f-değeri sırasıyla %99,18, %99,38 ve %99,28 olmuştur. Çapraz doğrulama (k=10) ile 0,001 std. değeri ile %99,8 başarım yakalanmıştır. Bu performans değerleri kategorik özellik yapısının kullanıldığı literatürdeki tüm çalışmalardan yüksektir. Elde edilen sonuçlar metin tabanlı özellikler barındıran suç faaliyet raporlarının kategorik olarak model eğitimlerinde kullanılması yerine Doc2Vec gibi doğal dil işleme teknikleri ile işlenebilir vektörlere dönüştürülmesinin sınıflandırma performansına doğrudan katkı sunabildiğini göstermiş ve daha az ön işlem sayesinde daha verimli bir modelin ortaya çıkmasını sağlamıştır.

Anahtar Kelimeler: Suç tahminlemesi, Kriminoloji, Doc2Vec, Topluluk öğrenme

1. INTRODUCTION

Crime is a set of actions and behaviors that damage the general fabric of society, endanger the safety of people's property and lives, and have a penal equivalent according to the law. Accordingly, criminal behavior among people living in a community is one of the natural consequences of being a society. Governments, state security forces, and scientists have important duties in the fight against crime. There are numerous research studies in this field. The main purpose of these studies is to investigate why people commit crimes, the mental and physical structures of people who commit crimes, the material reasons that push people to commit crimes, and to help the security organs to the extent possible. First, crime is a human act since its source is human. People have been living in communities since the earliest ages of history. In the society in which people live, there are inequalities in physical, sociological, and, as a result, psychological aspects. This situation prepares the ground for people to commit crimes. For this reason, committing a crime is an action that has been occurring since the early periods of history and has become more and more common with the increasing population. Among social problems, crime is perhaps the most complex and interesting one. In addition to the causes of the crime phenomenon, issues such as the rise and fall in crime rates and how crime is committed have been the subject of various studies [1].

Crime represents an important threat to humanity and is a concept that needs to be analyzed. Today, big cities attract people due to their wide range of opportunities, level of development, and work capacities. This situation brings about initiatives to facilitate people's lives in many areas, such as urban development, security, energy capacity, and environment. The most important of these initiatives is increasing the security capacity and strengthening the organizational structure. In this way, it will be possible to control the crime rates that determine the quality of life in cities and to create more livable cities. Many modern technological structures are used to ensure citizens' security in these cosmopolitan cities. It is still a big problem for today's cities, increasing daily. This problem is growing with the population [2]. It is the responsibility of security units to control and reduce this problem. However, to fulfill this responsibility, the crimes committed in cities should be analyzed in depth, and the threat levels to society should be determined. To make this analysis more detailed, country- and city-based crime statistics from the past to the present are prepared and shared with researchers [3-5]. In this way, it is possible to determine the types of crime, to reveal the regional densities of every kind of crime, and to conduct research to determine their prevalence. Regardless of the country, many crime records are reported daily in big cities, and many correspond to real crimes in the legal sense. In addition, some of these may be records that do not involve a crime. Making this distinction may be beneficial in terms of crime intervention.

The San Francisco Crime Dataset, shared for a competition to analyze this problem, is an open-access dataset containing information on more than 800,000 crime records from 39 different crime types. It includes details of all crime reports made to the police over 12 years. With this database, it is possible to analyze crimes, evaluate, and model for the future with past crime records. In this way, it is possible to obtain information about the general situation of the city, and it helps to make plans for good protection [6].

The main purpose of this study is to propose, test, and share the results of a machine learning-based methodology for the units responsible for the security of a city to use their capacities more efficiently and to better predict and fight against crime. In this way, it will be possible to detect environmental and regional crimes and to reveal possible scenarios based on past data. The most basic of these scenarios is evaluating whether the reported crime records are real crimes. Because not every crime record will contain a real crime, it is very critical to predict whether this crime report is a real criminal incident with the help of the reported region and basic information. This way, it will be possible to intervene only in criminal incidents, and non-criminal incidents will be automatically evaluated without any loss of labor thanks to this prediction. To perform this modeling, a series of analyses such as; (1) converting the San Francisco dataset into a trainable form with machine learning models, (2) analyzing the regional distribution of criminal and non-criminal incidents in the dataset, (3) revealing the variability of the crime in terms of date time and day must be completed, and all these processes have been carried out in this study. Then, in the modeling phase, steps related to machine learning models, such as data pre-processing, classifier selection, training and testing processes, and data balancing, were applied. Thus, detecting criminal and non-criminal cases with higher success was possible.

The main contributions of this work:

1- Since the features can be used without being categorized with the proposed model structure, all spatial and temporal information was used for classification.
2- Thanks to the doc2vec, unlike the literature, the data does not need to be categorized.
3- The proposed model differs from the literature's unique stacking ensemble model structure and 8 different classifier.
4- The 99% accuracy value obtained for 15 classes is higher than the studies in the literature.
5- A problem with 15 crime classes has been addressed for the first time in the literature and the problem was classified with higher performance than the 2-class problems.

2. LITERATURE REVIEW

Many studies have been conducted to develop future prediction models by using statistical data from past years to detect crime types. This section reviews and summarizes studies on crime analysis and classification in the literature. When the related literature is analyzed, more limited studies are found than most applied problems.

Junbo et. al. [7] comparatively examined the performance of different classifier models using the San Francisco dataset and analyzed the advantages and disadvantages of each model. The results indicate that, unlike other studies in the literature, Naive Bayes performs less than different classifiers. It has been stated that KNN and GB algorithms show higher performance, but their training and testing times are longer. Khan et. al. [8] compared the results of different classifiers for crime detection and category analysis. In the tests using the San Francisco dataset, the 10 categories with the highest number of crimes were selected, and the tests were carried out. The GB algorithm achieved the highest performance and obtained a 98.5% accuracy value. The results support other studies in the literature and prove that the GB algorithm can be more successful for this problem. Wu et. al. [9] surveyed crime prediction by analyzing the San Francisco dataset. Spatial and temporal crime data were analyzed, and each crime category's probabilities were calculated. In the evaluation of a 36-class problem, an accuracy of 28.51% was obtained. In the binary classification tests, an error value of 2.78% was obtained and it is stated that the police departments work efficiently by evaluating the occurrence or non-occurrence of crime. It has been proved that time and spatial data contain serious information about crime. Yehya et. al. [10] used coordinate data, address, date, and day of the crime to analyze and predict crimes. Tests were carried out for different machine-learning models with the San Francisco dataset used in this study. As a result of the tests, a RMSE value of 2.39 was obtained for the Random Forest classifier. Arslan et. al. [11] conducted a study in which 10 features were used depending on the time and location of the crime and classification was made with Random Forest. In the study, 86.5% success was achieved in the tests performed using the San Francisco dataset, and the AUC value was 0.98.

Aldossori et. al. [12] analyzed potential crime situations to enable law enforcement officers to detect crimes and suspicious situations. It aims to detect regional crime categories using existing crime records with machine learning. In the tests where the CLEAR dataset of the Chicago Police Department was used and nine features were extracted and given to Naïve Bayes and Decision Tree classifiers, 91.68% success was achieved. Djon et. al. [13] aimed to analyze only theft crimes among different types of crimes. The prediction was made by using the spatio-temporal information of the crime. As a result of data preprocessing and hyperparameter optimization, an f-score of 86% was obtained with XGB. Forradellas et. al. [14] proposed a regional crime prediction model for Buenos Aires. The model is based on modeling the records obtained for crimes, such as homicide and theft, between 2016 and 2019 with machine learning models. The average MAE and MSE values obtained with the proposed model after a series of preprocessing and very transformation processes were 0.4095 and 1.4602, respectively. Kim et. al. [15] analyzed Vancouver's crime records for the last 15 years. The data were obtained from the National Crime Records Bureau. Crime patterns were extracted and classified with various techniques using the data obtained. Among the structures using multiple models, the highest success rate was between 39% and 44% with KNN. Alves et. al. [16] studied crime prediction by examining the correlation structure between past crime records. Crime classification was performed with 97% success using the random forest-based model. While it was determined that urban structure is an issue that directly affects crime prediction and the main factor for Brazil is unemployment. This situation proves that models should be prepared using country and city-specific records.

Wu et. al. [17] aimed to predict the crime pattern and rate using data mining and machine learning based on regional historical crime records in YD Country. Using the records of 2012-2015 and the Bayesian network and Random Forest classifier models, the link between the type of crime and the job and gender of the offenders was estimated. As a result, it was observed that the Bayesian network has low correlation and poor performance in terms of various crimes and characteristics compared to other classifiers. Bandekar et. al. [18] proposed a structure based on machine learning models based on the location, time, and nature of crimes committed in India. They aim to reveal the connection between settlement locations and the type of crime. Thus, regional determination of risk factors and some predictions about crime were provided. As a result, location detection with Bayesian, Levenberg, and Scaled algorithm and statistical correlation with ANOVA achieved 78% success. Gül [19] examined the analysis and research methods used to prepare trend series related to crime and to make predictions on the development of crime and discussed the weaknesses and strengths of these methods, research results, and projections. In this study, it is examined important factors such as demographic variables, macroeconomic factors, technology, globalization, and new strategies in the fight against crime, which may impact the intensity and structure of crime. In the conclusion section, it is suggested future research and crime forecasts to be conducted in Turkey. Iqbal et. al. [20] proposed a crime category prediction model using naive Bayes and decision tree classifiers that can work for different states of the United States. The results of both classifiers were compared with the model, and the highest performance was obtained for the decision tree classifier, with an accuracy of 83.95%. Saeed et al. [21] used data mining in crime and community datasets of the United States, Naive Bayes, and Decision Tree for crime prediction and analysis and compared the results. As a result of this comparison, they claimed that, unlike Iqbal, Naive Bayes performs better than Decision Tree. Another study with the same dataset [22] conducted an experimental study in which they calculated the AUC metric and used different classifiers for crime prediction and analysis. They compared the results by integrating other feature selection methods into the model. As a result, the highest performance was obtained for the Naive Bayes classifier, and the AUC value was 0.898.

When the studies in Table 1 were examined, it was seen that there were limited studies on the use of artificial intelligence techniques in crime analysis and detection. It is thought that the main reason for this is that datasets on the subject are not common, and the datasets do not have definite value. In addition, it has been seen in existing studies that classification is made with basic machine-learning models. The use of deep learning or ensemble models is quite limited. In addition, it has been observed that categorizable features are selected and used in the training and testing processes. This study aims to use non-categorical features to solve these two basic constraints and to develop a model with high classification success with the ensemble model structure. In addition, it has been understood that regional, city, and country-based studies have been carried out on crime analysis. Modeling on crime analysis varies from country to country due to changes in socio-economic conditions. In this study, analyses were made, and tests were carried out with the data set of the city of San Francisco.

Table 1. Literature review

Paper and year	Dataset usage	Classifier design	Results
Junbo (2018)	San Francisco Crime Dataset	Naïve Bayes, KNN, GB	KNN: 97.58% NB: 97.40% GB: 97.60%
Khan (2022)	San Francisco Crime Dataset (with 10 categories with the highest number of crimes)	GB	98.75%
Wu (2016)	San Francisco Crime Dataset with Spatial and Temporal Data	KNN, LR	28.51% for 36 class 97.12% for binary
Yehya (2016)	San Francisco Crime Dataset with coordinate data, address, date and day of crime	RF	97.61%
Aldossori (2020)	CLEAR (Chicago Police Department with 9 features)	NB, DT	91.68%
Djon(2023)	Chicago Crime Dataset with spatio-temporal features for analyzing only theft crime	XGB	86%
Forradellas (2021)	Buenos Aires Crime Dataset (including 2016 and 2019 crime records)	K-means and neural network with 2 hidden layer	0.4095 (MAE) 1.4602 (MSE)
Kim (2018)	Vancouver crime dataset (15 years data)	KNN	44%
Alves (2018)	Brazil crime dataset (10 years data)	RF	97%
Bandekar (2020)	India crime dataset	Bayesian, Levenberg, Scaled algorithms, ANOVA	78% for crime location detection
Iqbal (2013)	United States dataset	Bayes, DT	83.95%
Saeed(2015)	USA Communities and Crime Unnormalized dataset	NB, DT and rule mining	80% to 95% (changing each crime type)
Shojaee (2013)	USA Communities and Crime Unnormalized dataset	Naïve Bayes	0.898 AUC
Arslan (2023)	San Fransisco Crime Dateset	Random Forest	86.5%

3. MATERIAL AND METHOD

Within the scope of this study, a model for crime category prediction was designed and tested. The flow chart for this model is shown in Figure 1. As the first stage of the model, the training dataset consists of San Francisco crime incidents retrieved from Kaggle [20]. To model in this published dataset, the features must be pre-processed and made trainable and vectorizable with the Doc2Vec model. To meet this requirement, a series of pre-processing stages such as stemming, lemmatization, tokenization, spelling correction, stop words removal, remove punctions, common words removal, and rare words removal, which are commonly preferred in natural language processing processes, have been tried for testing purposes. Only stop word removal, lemmatization, and stemming operations were applied to the data because they contributed positively to the classification. From the data resulting from this application, the 15 categories with the most records were selected and divided into 70% training and 30% testing. After this separation, the training and test data were trained separately with the Doc2Vec network and converted into vectors. A standardization process was applied to the data resulting from this transformation and was given to the stacking ensemble model, which includes different Machine Learning models. The model structure, which completed the learning phase with training data, was tested with test data, and its performance was measured according to different evaluation metrics for this problem with 15 classes. The main purpose of this model structure is to enable classification without converting the features into categorical ones, unlike the literature, and, as a result, to achieve high classification performance.

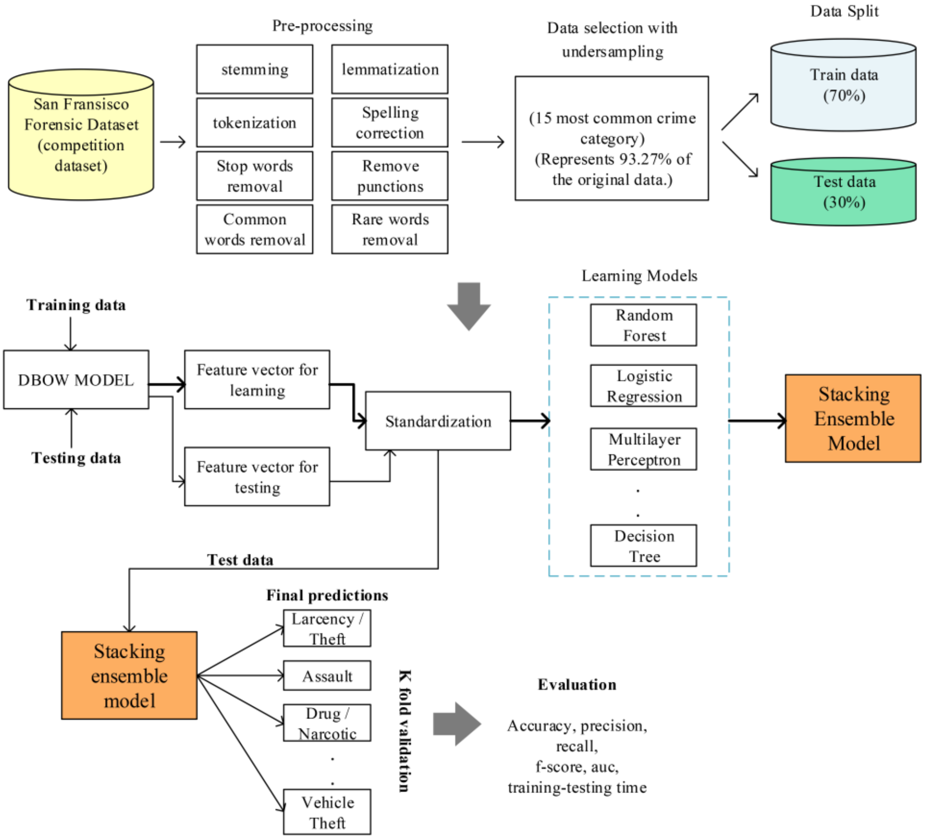

Figure 1. Methodology diagram of proposed model

3.1. Dataset Details

Learning models have been proposed recently for detecting and preventing crimes. It aims to predict crime types using these models and assist law enforcement with these predictions. Many datasets are openly distributed for this purpose. Some of these also attract attention as competition datasets. However, the main difference between these datasets is that each belongs to a different city and country. Their features are similar, but the records belong to different cities. In this study, datasets were examined, and it was desired to determine a dataset that was the most recent, collected in the widest year range, and had many samples. In addition, since a text-based classification and analysis will be carried out in this study, the presence of features that cannot be made categorical was an important criterion in selection. In line with these criteria, The San Francisco Dataset, which contains crime records of San Francisco, was selected because it is a competition dataset, contains records distributed over 12 years, and contains text-based data such as Address, PDDistrict, and Descript.

The San Francisco dataset contains 878 thousand crime records obtained from annual crime reports for 12 years between 2003 and 2015. There are 39 different crime types in this data set, with various numbers of crime records for each crime type and some have fewer than 1000 records. Within the scope of this study, the dataset was filtered for crime types with 10000 or more records to balance the number of samples between crime types and prevent the model from converging to any one class. As a result of this filtering, a problem with 15 classes, including 14 different criminal category and one non-criminal category, emerged.

The original record numbers for each class are shown in Table 2. When the original numbers were examined, it was seen that there was a serious imbalancing between crime categories. When model training is performed with this imbalanced data set, the models tend to be biased against the majority class. As a result, while the majority classes are successful in prediction, the minority class may perform poorly. Thus, overfitting to certain types can cause generalization problems. When data is increased with over-sampling to solve this problem, the model may memorize the data and lead to overfitting. For all these reasons, balancing was made with data under-sampling. After that, the training and testing processes of the proposed model were carried out with a total of 280 thousand records, approximately 10 thousand records for each crime category.

Table 2. Dataset sample distribution for different types of offences and non-criminal offences

Crime type	**Orjinal**	**After undersampling with "minority class"**
LARCENY/THEFT	174900	10000
OTHER OFFENSES	126206	10000
ASSAULT	94525	10000
DRUG/NARCOTIC	53971	10000
VEHICLE THEFT	53779	10000
VANDALISM	78793	10000
WARRANTS	42212	10000
BURGLARY	75711	10000
SUSPICIOUS OCC	53694	10000
MISSING PERSON	25989	10000
ROBBERY	45503	10000
FRAUD	20974	10000
FORGERY/ COUNTERFEITING	12293	10000
SECONDARY CODES	10000	10000
NON-CRIMINAL OFFENCES	139975	139975

For each record in the dataset, Dates (timestamp of crime), Descript (short description of crime), DayofWeek, PdDistrict (district of the city), Resolution (brief description of crime resolution), Address (Address of the crime), X (longitude) and Y (Latitude). In the first stage, these features were divided into sub-features, as shown in Table 3.

Table 3. Feature conversion from original dataset to proposed model

Original feature from dataset	Sub-category conversion
Dates Sample: (2015-05-13 23:53:00)	Year, Month, Day, Hour, Minutes, Seconds
Descript	Descript
Dayofweek Sample: (Wednesday)	Day of week
PdDistrict	PdDistrict
Resolution	Resolution
Address	Address
X	X
Y	Y

As a result of this segmentation, feature selection was made on 11 features that emerged except Address and Description.

There are many different approaches to feature selection [23]. Among these, the SHAP (SHapley Additive exPlanations) [24] feature selection method was used in this study. Shapley distributes the values more fairly than methods such as LIME and the difference between the estimate and the average estimate, and proves this by basing it on a theory. In addition, it enables comparative evaluation of each feature on a class basis.

SHAP value was calculated for 11 features obtained as a result of preprocessing for 280 thousand records, and classification contribution levels were calculated. The outcomes are given in Figure 2. The larger Shapley value indicates that the feature is more of a classifier. As shown in Figure 2, the second value does not contribute positively to any criminal or non-criminal classification classes. The main reason is that it has the same value for all data.

The "resolution" feature contains textual data with clues to the crime. This situation reduces the objectivity of the proposed model. For this reason, although "Resolution" has the discriminative ability, as shown in Figure 2, it was removed from the dataset. The remaining 9 features and "Address" and "Description" features were merged and converted into a single textual feature. The vector resulting from this transformation was used in the training and testing processes of the Doc2Vec model. The main purpose of this two-layer long pre-processing process is to both reveal the most distinctive features and use them in classification by converting text-based features into vectors. This way, higher performance was achieved compared to categorical-based models, as given in this study.

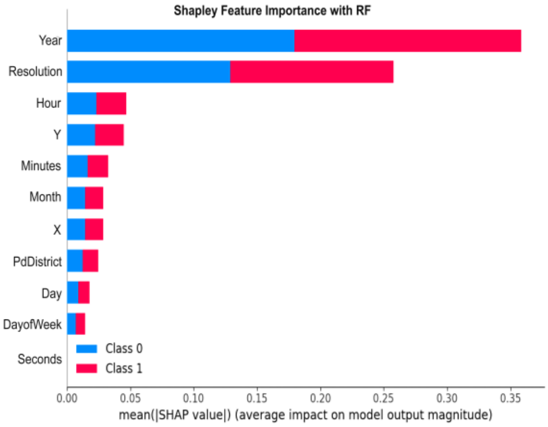

Figure 2. Feature importance plot for random forest according to Shap value (Class0: criminal, Class1: non-criminal)

3.2. Classifier Selection for Ensemble Model

In this study, for the 1st and 2nd layers to design the ensemble model, 8 different classifiers were used, namely Random Forest, Logistic Regression, Multilayer Perceptron, Support Vector Classifier, K-nearest neighbor, XGB, Gaussian Naïve Bayes, and Decision Tree, which have different architectural structures and are widely used to solve various problems [25-27]. After this, each classifier was tested independently of each other. During this testing process, hyperparameter optimization was performed. Therefore, the performance values in the next section show the highest achievements obtained due to optimization. To place 8 different classifiers in the stacking ensemble model, selecting the machine learning model to work as a decision-maker in the 1st and 2nd Layers is necessary. This selection was made according to the pre-test results, and the SVC classifier with the highest result was used as the decision-maker in the model.

3.3. DBOW Model

The DBOW (distributed bag of words) model is a variant of a popular word embedding method called Word2Vec [28] and the architecture of model is shown in Figure 3. Word2Vec is used to learn word vectors (numerical representations) and is often used in natural language processing (NLP) tasks. The DBOW model is one of the two main variants of Word2Vec [29, 30]. The working principle of the DBOW model is as follows:

1. Data Preparation: The first step is to create a vocabulary representing a large text dataset. Each word is assigned a unique vector.
2. Training: The DBOW model learns word vectors without considering the context of a word. It treats each word independently and tries to predict a random word within a "context window".
3. Prediction: Without capturing each word's context in the dataset, the model makes predictions using other words around that word. These predictions are used to capture word associations in the dataset.
4. Training Result: The model improves by comparing its predictions with the actual labels. This is used to update the word vectors at each step.

As a result, the DBOW model treats each word independently and learns word vectors instead of directly modeling word relations within the text. These vectors represent the word's semantic meaning and its position in the text. The DBOW model can be used in many NLP tasks, such as word similarity measurement, document classification, and text generation.

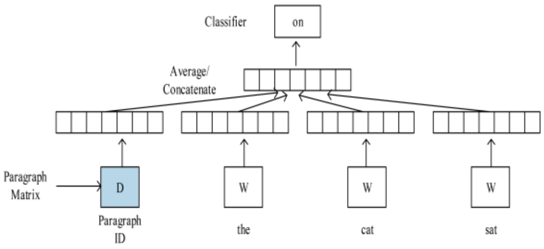

Figure 3. DBOW model architecture [28]

3.4. Stacking Ensemble Model Structure (Proposed Model)

Stacking ensemble model is an ensemble technique that aims to create a more powerful predictor by combining multiple machine learning models [31]. The basic components of the stacking ensemble model are as follows:

1. Base Models: The first step is to create multiple base models using different machine learning algorithms or the same algorithm with other parameters. For example, algorithms such as decision trees, random forests, support vector machines (SVM), or gradient boosting models can be used.
2. Meta Model (Stacking Model): The second step is to create a “meta-model” or “top model” using the predictions of the base models. This meta-model takes the projections of the base models as input and makes final predictions using these predictions. Usually, this meta-model can be a regression or a classification model because its purpose is to combine the projections of the base models and obtain a better result.
3. Training and Evaluation: The base models are trained separately, and then the meta-model is trained using the predictions of these models. The stacking model can be optimized, and its performance can be evaluated by techniques such as cross-validation.
4. Combining Predictions: In the last step, the stacking model is used to make predictions on new data. The predictions of the base models are given as input to the meta-model, and the meta-model uses these predictions to produce the final predictions.

The stacking ensemble model structure proposed in this study is shown in Figure-4. Stacking ensemble models allows different models to balance each other's shortcomings and achieve a stronger performance.

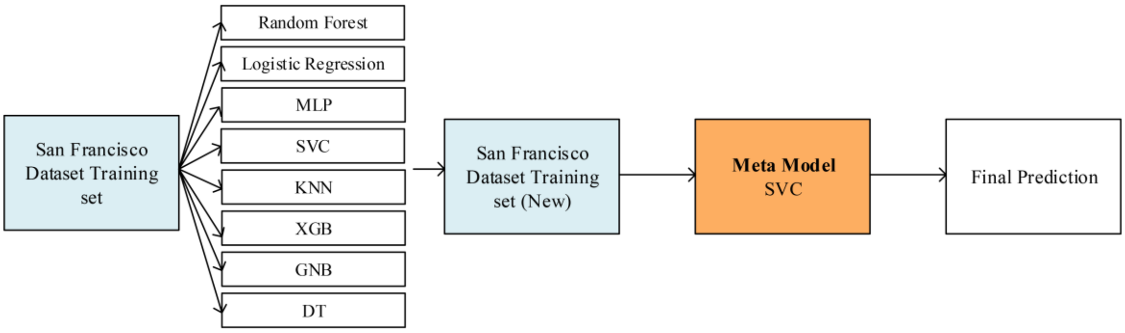

Figure 4. Stacking ensemble model structure

3.5. Evaluation Metrics

Evaluation metrics used to measure, analyze, or compare the performance of a model. These metrics determine how good or bad a model is and are usually based on comparing a model's predictions with actual values. The results of this study were measured with the accuracy, precision, recall, f-score and AUC value given in Equations 1-4 and were evaluated comparatively in the next section. For multiclass classification problems, TP, TN, FP, and FN numbers are found for each class, and metrics are calculated. In this study, performance measurement was carried out using the equations below for the 15-class dataset.

1. Precision: Precision measures how many of the samples predicted as positive are actually positive. It aims to reduce the number of false positives.

$$Precision\ =\ \frac{\sum_{i=1}^{15} TP_i}{\sum_{i=1}^{15}(TP_i+FP_i)} \tag{1}$$

2. Sensitivity (Recall): Sensitivity measures how many true positives are correctly predicted. It aims to reduce the number of false negatives.

$$Recall\ =\ \frac{\sum_{i=1}^{15} TP_i}{\sum_{i=1}^{15}(TP_i+FN_i)} \tag{2}$$

3. F-Score: F-Score is a metric that represents the balance between precision and sensitivity. It summarises the classification performance of the model.

$$F - score\ =\ \frac{2*Precision*Recall}{Precision+Recall} \tag{3}$$

4. Accuracy: Accuracy is the ratio of correct predictions to total predictions. It is generally used in classification problems.

$$Accuracy\ =\ \frac{\sum_{i=1}^{15}(TP_i+TN_i)}{\sum_{i=1}^{15}(TP_i+TN_i+FP_i+FN_i)} \tag{4}$$

5. ROC Curve and AUC: The ROC curve shows the relationship between sensitivity and specificity at different limiting thresholds. AUC measures the area under the ROC curve and evaluates the classification ability of the model.

Evaluation metrics are used to measure the model's success and improve the model quantitatively. Which metrics to use may vary depending on the data type, task, and objective. An evaluation metric is an important tool for understanding the model's performance and tuning the model.

4. TEST RESULTS

This study proposes a novel model for classifying criminal activities using a Doc2Vec structure that can produce a numerical representation of texts regardless of length and a stacking ensemble model with 11 different machine-learning models.

Using data pre-processing techniques on the San Francisco Crime Classification datasets, machine learning algorithms were used to analyze the data. Predictive classification was performed with the features extracted from the new datasets obtained using data augmentation techniques. All results obtained for the San Francisco Crime Classification data are listed in Table 4.

Table 4. Classification results for different ML algorithms

Classifiers	Accuracy (%)	Precision (%)	Recall (%)	Fscore (%)
Logistic Regression	98.83	98.83	98.83	98.83
Random Forest	98.12	98.13	98.12	98.12
Decision Tree	84.49	84.53	84.49	84.5
Gaussian NB	98.14	98.18	98.14	98.15
Linear Discriminant Analysis	98.64	98.64	98.64	98.64
Ada Boost	89.58	89.78	89.58	89.65
Extra Trees	98.69	98.69	98.69	98.69
SVC	98.89	98.89	98.89	98.89
KNeighbors	98.75	98.75	98.75	98.75
XGB	94.14	94.18	94.14	94.15
MLP	97.24	97.25	97.24	97.24
Voting Classifier	99.11	99.11	99.11	99.11
Stacking Classifier (Proposed Model)	**99.28**	**99.18**	**99.38**	**99.2**

When Table 3 is analyzed, it is seen that all classifiers generally give high performance with the feature vector extracted with Doc2Vec. Accuracy value varies between 89.58 and 99.31. Although there are normal values above 90%, the lowest value is the DecisionTreeClassifier model, with 84.49%. Precision, recall, and f-score values have similar values with accuracy. The Stacking Classifier model proposed in this study obtained the highest performance. According to the literature, achieving the highest performance for a model with 15 classes is an important output. The closest to the model we studied is the VotingClassifier model, with 99.11%. The critical point is that the features are extracted as text-based rather than definite and valuable information, such as address, which is included in the classification model. To verify these results with cross-validation, the results of 10-fold testing of all classifiers are given in Figure 5.

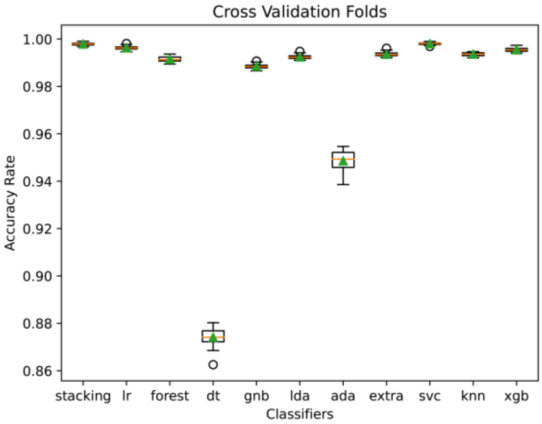

Figure 5. Cross validation results of ML models and proposed model

Figure 5 shows the results obtained by machine learning algorithms after 10-fold cross-validation on the San Francisco Crime Classification dataset. The graph can be interpreted based on accuracy value. In this case, the classifier with the highest accuracy value is the stacking ensemble, and the median value is 99.8%. Such graphs help us to analyze the statistical properties of the data. It allows us to comment on the accuracy value's central tendency, dispersion, and outliers. While the spread of the Ada classifier is the largest, the spread of the Stacking ensemble classifier is the smallest. In this case, the most homogeneous distribution in terms of accuracy value is in the Stacking ensemble classifier, while the most heterogeneous distribution is in the Ada classifier. Logistic regression, Decision Tree, and Extra tree classifiers have outliers. Outliers are the results that are outside the calculated minimum and maximum value. In general, when we look at the Q1 (first quartile), IQR (interquartile range), and Q3 (third quartile) values of all classifiers, it is also seen that they show a symmetrical distribution rather than a skewed distribution. The lowest accuracy value was observed in the Decision Tree classifier. The median (average) accuracy values are close to each other in logistic regression, SVM, and stacking ensemble classifiers. The orange line in the graph is the median value. It shows the central tendency of the accuracy value.

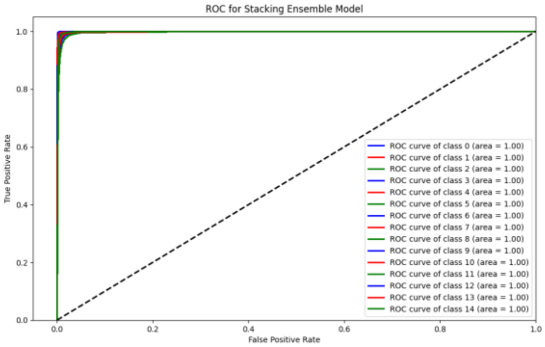

Figure 6. ROC for stacking ensemble model

Figure 6 shows the ROC Curve of the Stacking Ensemble Model. If we interpret the result using this graph, AUC value 1 was obtained in 15 classes. This indicates that the model has very good classification ability. This graph's upper left corner is where false positive and false negative predictions are minimized. In this region, balance is achieved for all classes in the diagram. In addition, 10000 samples were worked within each of the 15 classes and were equally distributed. There is no imbalance between classes. The fact that the graph curves of the 15 classes are similar shows that the model predicts all classes equally. The dashed line in the graph shows the random prediction of the model. Since the curves for 15 classes are located on this diagonal, it is clear that the classifier performs better than random predictions.

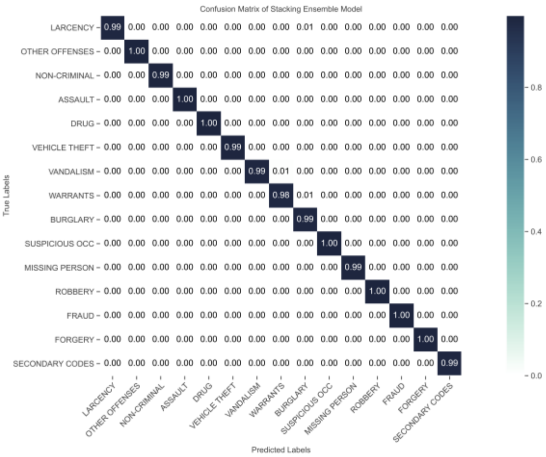

Figure 7. Confusion matrix of the stacking ensemble model

The confusion matrix for the stacking ensemble model is shown in Figure 7. When the results are analyzed according to this figure, the classes with the highest accuracy are assault, drugs, suspicious incidents, robbery, fraud, and forgery, with a 100% recognition rate. The class with the lowest recognition rate is warranted, with an accuracy rate of 98%. The overall success rate is 99%. In addition, 98% and above accuracy values were obtained in 15 classes.

5. CONCLUSION

The tests conducted in this study aim to determine the criminal and non-criminal status of crime reports in crime prediction. San Francisco Crime Classification, which includes crimes committed for 12 years, was used as a data set. Tests for 11 different machine learning models were performed. Accordingly, an accuracy value of 99.28% was obtained for the 15 offense categories with the highest crime records, while the precision, recall, and f-score values were 99.18%, 99.38%, and 99.20%, respectively. With cross-validation (k=10), 99.8% success was achieved with a std. Value of 0.001. These performance values are higher than those of all the studies in the literature using categorical feature structures. The obtained results show that the transformation of criminal activity reports, which contain text-based features, into vectors that can be processed with natural language processing techniques such as Doc2Vec instead of using them categorically in model training can directly contribute to classification performance and provides a more efficient model with less pre-processing. In addition, the proposed model can help security organizations to develop a more effective crime response system. These results prove that crimes can be assessed spatially for cities and that the probability of crime depends on regional conditions.

In the future, the same classification model can be tested on datasets from different cities and countries and evaluated whether it has similar performances. In addition, studies can continue transforming more crime-related data into features through feature engineering. Increasing the number of crime types with a mixture of temporal and spatial analyses can make more temporal analyses of crime using time series. Studies can perform better using more complex and advanced classification models.

6. REFERENCES

1. İçli, T.G., 1993. Türkiye'de Suçlular (Sosyal Kültürel ve Ekonomik Özellikleri. Atatürk Kültür, Dil ve Tarih Kurumu Atatürk Kültür Merkezi Yayını, Ankara, 71.

2. Hochstetler, J., Hochstetler, L., Fu, S., 2016. An Optimal Police Patrol Planning Strategy for Smart City Safety. IEEE 18th International Conference on High Performance Computing and Communications, Sydney, Australia, 1256-1263.

3. Open Government, https://www.data.gov/open-gov/, Access date: Haziran 2023.

4. Data.world Crime Datasets, https://data.world/datasets/crime, Access date: Temmuz 2023.

5. All Data Related to Crime And Justice, https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/datalist?filter=datasets, Access date: Ağustos 2023.

6. Pradhan, I., Potika, K., Eirinaki, M., Potikas, P., 2019. Exploratory Data Analysis and Crime Prediction for Smart Cities. Proceedings of the 23rd International Database Applications and Engineering Symposium on - IDEAS '19, Athens, Greece, 1-9.

7. Ke, J., Li, X., Chen, J., 2018. San Fransisco Crime Classification (Report), Jocobs School of Engineering, San Diego, 7.

8. Khan, M., Ali, A., Alharbi, Y., 2022. Predicting and Preventing Crime: A Crime Prediction Model using San Francisco Crime Data by Classification Techniques. Complexity, 1-13.

9. Wu, X., 2016. An Informative and Predictive Analysis of the San Francisco Police Department Crime Data. M.Sc., University of California, Los Angeles, 11.

10. Abouelnaga, Y., 2016. San Francisco Crime Classification", arXiv:1607.03626.

11. Arslan, R.S., Dülgeroğlu, B., 2023. Crime Classification using Categorical Feature Engineering and Machine Learning. International Ankara Congress on Multidisciplinary Studies-VI, Ankara, Turkey, 1-8.

12. Aldossari, B.S., Alqahtani, F.M., Alshahrani, N.S., Alhammam, M.M., Alzamanan, R.M., Aslam, N.I., 2020. A Comparative Study of Decision Tree and Naive Bayes Machine Learning Model for Crime Category Prediction in Chicago. Proceedings of 2020 6th International Conference on Computing and Data Engineering, Sanya, China, 34-38.

13. Deborah, D., Jitesh, J., Kieron, D., Vincent, T., 2023. A Comparative Analysis of Multiple Methods for Predicting a Specific Type of Crime in the City of Chicago. ArXiv: 2304.13464.

14. Reier Forradellas, R.F., Náñez Alonso, S.L., Jorge-Vazquez, J., Rodriguez, M.L., 2020. Applied Machine Learning in Social Sciences: Neural Networks and Crime Prediction. Social Sciences, 10(1), 4.

15. Kim, S., Joshi, P., Kalsi, P.S., Taheri, P., 2018. Crime Analysis through Machine Learning. 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, Canada, 1-6.

16. Alves, L.G.A., Ribeiro, H.V., Rodrigues, F.A., 2018. Crime Prediction through Urban Metrics and Statistical Learning. Physica A: Statistical Mechanics and Its Applications, 505, 435-443.

17. Wu, S., Wang, C., Cao, H., Jia, X., 2020. Crime Prediction using Data Mining and Machine Learning. Intell. Syst. Comput., Springer Verlag, 905, 360-375.

18. Bandekar, S.R., Vijayalakshmi, C., 2020. Design and Analysis of Machine Learning algorithms for the Reduction of Crime Rates in India. Procedia Computer Science, 172, 122-127.

19. Gül, S., Polat, A., 2009. Kamu Güvenlik Politikalarının Oluşturulmasında Yeni Bir Yaklaşım: Suç Tahmini. Türk İdare Dergisi. 463 (81), 131-156.

20. Iqbal, R., 2013. An Experimental Study of Classification Algorithms for Crime Prediction. Indian Journal of Science and Technology, 6(3), 1-7.

21. Saeed, U., Sarim, M., Usmani, A., Mukhtar, A., Basit, S.A., Kashif Riffat, S., 2015. Application of Machine Learning Algorithms in Crime Classification and Classification Rule Mining. Research Journal of Recent Sciences, 4(3), 106-114.

22. Shojaee, S., Mustapha, A., Fatimah, S., Jabar, A., 2013. A Study on Classification Learning Algorithms to Predict Crime Status. International Journal of Digital Content Technology and its Applications, 7(9), 361-371.

23. Arslan, R.S., 2021. Comparison of Feature Selection Methods in Security Analysis of Android. 2021 6th International Conference on Computer Science and Engineering (UBMK). Ankara, Turkey, 1-5.

24. Lundberg, S., Lee, S., 2017. A Unified Approach to Interpreting Model Predictions. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 1-11.

25. Hizlisoy, S., Arslan, R.S., 2021. Text Independent Speaker Recognition Based on MFCC and Machine Learning. Selcuk University Journal of Engineering Sciences, 20(03), 073-078.

26. Hizlisoy, S., Tüfekci, Z., 2020. Türkçe Müzikten Duygu Tanıma. European Journal of Science and Technology, Special Issue, 6-12.

27. Arslan, R.S., Yurttakal, A.H., 2020. K-Nearest Neighbour Classifier Usage for Permission based Malware Detection in Android. Icontech International Journal, 4(2), 15-27.

28. Quoc, L., Tomas, M., 2014. Distributed Representations of Sentences and Documents. Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2), 1188-1196.

29. Arslan, R.S., 2021. Kötücül Web Sayfalarının Tespitinde Doc2Vec Modeli ve Makine Öğrenmesi Yaklaşımı. European Journal of Science and Technology, 27, 792-801.

30. Arslan, R.S., 2021. Kötücül URL Filtreleme için Derin Öğrenme Modeli Tasarımı. European Journal of Science and Technology, 29, 122-128.

31. Arslan, R.S., 2021. Identify type of Android malware with Machine Learning Based Ensemble model. 2021 5th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Ankara, Turkey, 1-5.